The Accuracy and Completeness of Data Collected by Prospective and Retrospective Methods

J. Tobias Nagurney, MD, MPH, David F. M. Brown, MD, Swati Sane, MBBS, MPH, Justin B. Weiner, BA, Andrew C. Wang, Yuchiao Chang, PhD

Abstract

Objectives: To describe and test a model that compares the accuracy of data gathered prospectively versus retrospectively among adult emergency department patients admitted with chest pain. **Methods:** The authors developed a model of information flow from subject to medical record to the clinical study case report form, based on a literature review. To test this model, a bidirectional (prospective and retrospective) study was conducted, enrolling all eligible adult patients who were admitted with a chief complaint of chest pain. The authors interviewed patients in the emergency department to determine their chest pain history and established a prospective database; this was considered the criterion standard. Then, patient medical records were reviewed to determine the accuracy and completeness of the information available through a retrospective medical record review. **Results:** The model described applies the concepts of reliability and validity to information passed on by the study subject, the clinician, and the medical record abstractor. This study was comprised of 104 subjects, of which 63% were men and the median age was 63 years. Subjects were uncertain of responses for 0–8% of questions and responded differently upon reinterview for subsets of questions 0–30% of the time. The sensitivity of the medical record for risk factors for coronary artery disease was 0.77 to 0.93. Among the 88 subjects (85%) who indicated that their chest pain was substernal or left chest, the medical record described this location in 44%. Timing of the chest pain was the most difficult item to accurately capture from the medical record. **Conclusions:** Information obtained retrospectively from the abstraction of medical records is measurably less accurate than information obtained prospectively from research subjects. For certain items, more than half of the information is not available. This loss of information is related to the data types included in the study and by the assumptions that a researcher performing a retrospective study makes about implied versus explicitly stated responses. A model of information flow that incorporates the concepts of reliability and validity can be used to measure some of the loss of information that occurs at each step along the way from subject to clinician to medical record abstractor. **Key words:** data collection; retrospective; prospective; bias; precision; methodology; myocardial infarction. ACADEMIC EMERGENCY MEDICINE 2005; 12:884–895.

The quality of retrospective versus prospective research studies has been the long-standing subject of many articles in the epidemiologic and clinical literature.[1] It is generally believed that historical data gathered prospectively by direct subject interview is more complete and more accurate than data gathered retrospectively. In fact, studies have actually examined, by specialty, the relative distribution of both types of studies in the literature of that specialty.[2] In particular, retrospective data, usually obtained through medical record review, is fraught with the problems of missing data, conflicting data, and illegibility.[3] Interobserver reliability or reproducibility in the medical record abstraction process may affect precision and hence accuracy.[3–5] In addition to data degradation from lack of reliability, nonrandom bias may occur at any step where information is obtained or recorded.[6,7] This bias on the part of the subject, the clinician, or the researcher conducting the abstract of the medical record may affect validity.[8–10] While this qualitative belief regarding the comparative quality of prospectively versus retrospectively acquired data is widely held, few studies have attempted to quantify the differences between these two methods of data collection.

Chest pain is a common presenting complaint among emergency department (ED) patients. The sizeable literature that addresses chest pain usually includes a series of historical items generally accepted to be important.[11,12] Although many of the studies in this literature gathered data prospectively, many others gathered it retrospectively through medical record

From the Department of Emergency Medicine (JTN, DFMB, SS) and Department of Medicine, General Medicine Division (YC), Massachusetts General Hospital, Boston, MA; Division of Emergency Medicine, Harvard Medical School (JTN, DFMB, SS), Boston, MA; University of Michigan School of Medicine (JBW), Ann Arbor, MI; and Princeton University (ACW), Princeton, NJ.

Received August 21, 2004; revision received April 13, 2005; accepted April 22, 2005.

Presented as an oral presentation at the SAEM annual meeting, Boston, MA, May 2003.

Address for correspondence and reprints: J. Tobias Nagurney, MD, MPH, Department of Emergency Medicine, Clinics 115, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02115. Fax: 617-724-0917; e-mail: nagurney.john@mgh.harvard.edu.

doi:10.1197/j.aem.2005.04.021

review.[13] The accuracy of data obtained retrospectively through such a medical record review, as compared with prospectively acquired data, has rarely been measured. In those few cases in which it has, it is far less complete and accurate than information obtained directly from the subject prospectively.[14,15]

Our goal was to describe a model to characterize the process of data collection and to use this model to explain the loss of information between prospective and retrospective studies. Our secondary goal was to quantify the misclassification of information along steps of the information flow, using data acquired both prospectively and retrospectively from subjects with chest pain.

METHODS

Study Design. We conducted a bidirectional descriptive study. Our institutional review board approved this study. Informed consent was obtained from all participants before study enrollment.

Conceptual Model: The Process of Data Collection. We developed a conceptual model of information flow from a subject in a research study to both prospective and retrospective databases, based on a review of the clinical and epidemiologic literature. In establishing the model, we searched the reference libraries among titles of epidemiology, research methods in medicine, social relations, and health surveys. We enlisted the aid of a health survey specialist and searched textbooks and authors suggested by her.[6,9,16–20] Finally, we conducted an Ovid search (Ovid MEDLINE, English language, 1966–2002) with the help of a professional research librarian, searching on the key words of epidemiology, bias, recall, recall bias, mental recall, psychometrics, retrospective, prospective, research design, agreement, medical record, and self-report.

The model resulting from this literature search is shown in Figure 1. It begins with the subject's true medical history, as described by the subject. For prospective studies, the subject is interviewed directly by a researcher and data are entered directly into a prospective data set. In retrospective studies, this information is channeled through a clinician who obtains and records it in the medical record. Researchers then review the medical record and record information into a retrospective database. In a perfect world, the information obtained by either of these processes would be equivalent. In the real world, they are usually not because of data degradation, bias, or both. Moreover, this data degradation may occur among all three participants who play a role in the information flow: the subject, the clinician, or the researcher.

Table 1 describes standard definitions of the concepts involved in this model. Note that each of the three participants in the flow of information can represent a source of unreliability, bias, or both.

Methodology of Bidirectional Study. Historical data were gathered prospectively by subject interview using trained research assistants. The same data were obtained retrospectively by medical record review. To interpret potential disagreements between these two data sources, we interviewed the providers as well. For those historical items where the subject interview and the medical record disagreed, we used the information obtained from this provider interview to interpret discrepancies. While imperfect, this allowed us to estimate whether data were lost between subject and provider or between provider and medical record.

Study Setting and Population. This study was conducted at an urban university hospital ED with 73,000 annual visits. Care for patients presenting with chest pain is provided by residents from both internal medicine and emergency medicine under the supervision of emergency medicine faculty. All patients are seen by a resident and by a faculty member. Residents generally write complete ED histories and physical examinations, whereas faculty members write briefer notes. For this reason, we studied the ED note written by the resident. In our institution, ED notes are open-ended notes without any standardized templates.

All adult patients (21 years or older) who presented to the ED with chest pain as the chief complaint and who were admitted to the hospital to rule out myocardial infarction (MI) were screened. The intake period for the study was July 10, 2001, to August 31, 2001, weekdays from 9 AM to 5 PM. Our goal was to recruit approximately 100 patients so that the width of the 95% confidence interval (CI) of any estimates would be limited to within 10%. Patients were excluded if they were in acute distress or were believed to be incapable of providing informed consent for other reasons.

Study Protocol. Each subject was interviewed by means of a closed-ended questionnaire. Interviews were conducted by research assistants who read a questionnaire. They were trained by the principal investigator and were routinely monitored to assure the consistency and quality of the interviewing process. The same research assistants then interviewed the provider within one hour using the same questionnaire. The subject and provider questionnaires contained three sections, each including questions typically asked of patients with chest pain: the history of coronary artery disease (CAD) (two questions), the history of risk factors for CAD (three questions), and the details of the chest pain story (seven questions). This subject interview was considered the database that would have been obtained in a prospective study and was considered the criterion standard. To test the reproducibility of this history as given by subjects, a 20% subsample was interviewed again by the same research assistant within two hours of the first interview.

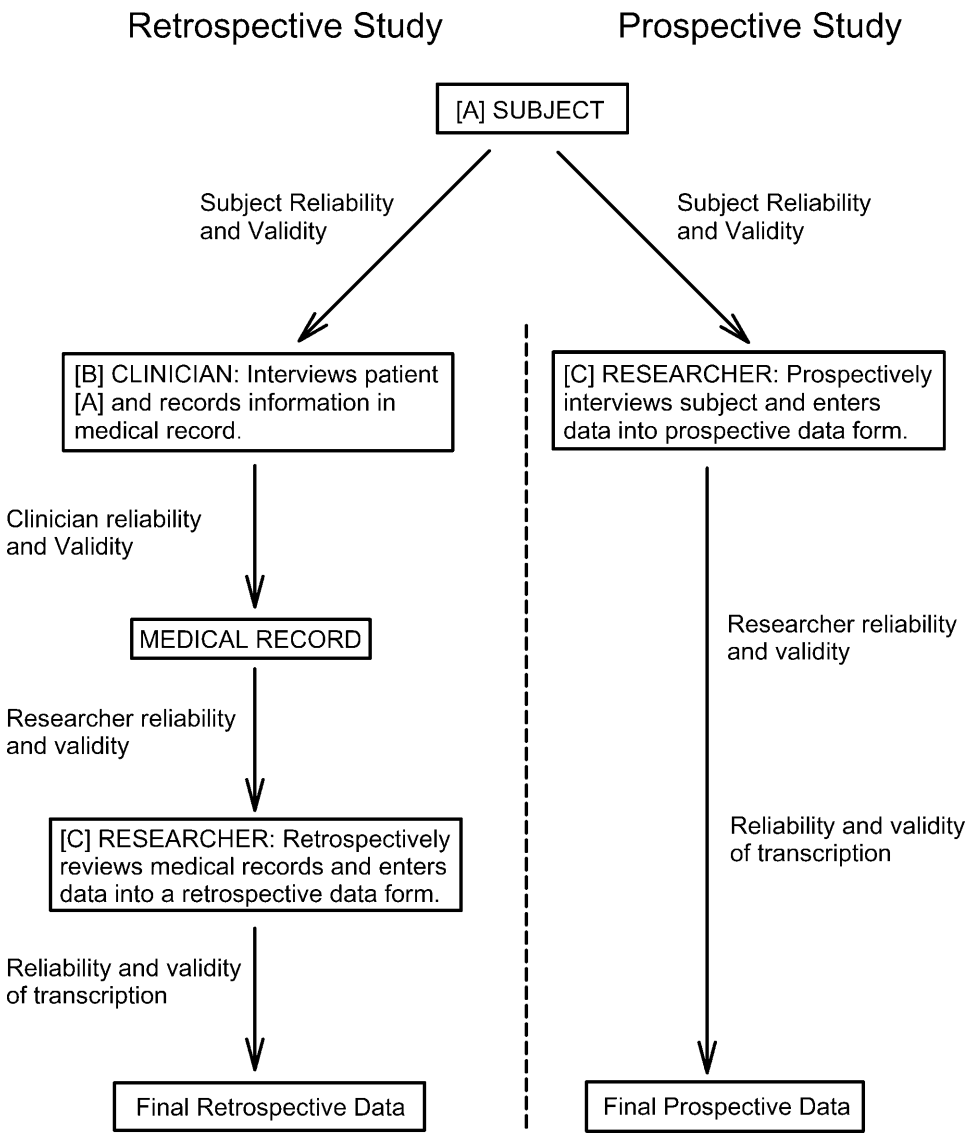

Figure 1. Conceptualization of information flow.

Subject Interview. Subjects were asked whether they had a history of CAD, a coronary artery bypass graft, hypertension, diabetes mellitus, or an elevated cholesterol level. These comorbidities were considered positive if the subject had been treated with medications for that illness within the previous six months. For all five of these variables, the subject responses were coded as yes, no, or uncertain. For chest pain location and quality, subjects were given five and six choices, respectively, and asked to choose the one that most closely matched their symptom. For chest pain timing, subjects were asked whether their chest pain was steady or intermittent and the time of onset of the most recent episode. For the variables of chest pain radiation, associated symptoms, and relieving factors, subjects were given five choices (including "other") for each historical item and asked to identify all that were present. For these last three variables, potential responses not specifically identified as present by the subject were considered negative. Historical items of which the subjects were themselves uncertain were quantified but excluded from further analysis, because the medical record could not be expected to accurately contain them. Hence, the sample size differs slightly for different variables.

Provider Interviews. Providers were told that they were being interviewed to determine information flow, but not the particular details of the study, to minimize a potential Hawthorne effect. The providers were asked the same questions as subjects in the same order. They were also provided the same definitions and criteria for variables. Responses were coded the same as for the subjects, namely yes, no, or uncertain.

Medical Record Abstraction Process. The medical record abstraction tool was in a form that paralleled the subject and provider questionnaires. For each

TABLE 1. Definitions

Truth	What really exists in the universe in relationship to the subject. Often not knowable. Researchers use what subjects tell them as an approximation of this truth. One can sometimes cross-reference with outside sources such as medical records.
Reliability	Consistency over repeated measurements. Also called reproducibility, precision, or internal consistency. It is related to the testing instrument or observer in a random manner. It is typically measurable.
Validity	A measure of how well the result measures the data item it is designed to measure as corroborated by outside criteria. Also called external consistency. Lack of validity is called bias and is usually nonrandom. Bias is difficult to measure and often goes undetected.
Subject reliability	The ability of the subject to tell and retell his or her story over and over in the same way. A measure of reproducibility in history telling.
Subject validity	The ability of the subject to tell his or her story without bias or distortion. Bias may enter for many reasons and from many sources. Examples of sources of bias include censoring, recall bias, and social desirability bias.
Clinician reliability	The ability of the clinician to hear and record subject information in a reproducible manner. Nonreproducibility at this step may occur at either the information gathering stage or recording stage.
Clinician validity	The ability of the clinician to hear and record subject information without bias. Bias may include censoring, which can affect either what the clinician thinks he or she hears or what he or she chooses to record.
Researcher reliability	The ability of the researcher to read the medical record and to record subject information in a reproducible manner. Sources of data degradation include missing information, conflicting data, and illegibility.
Researcher validity	The ability of the researcher to read the medical record and to record subject information without bias.

historical item, the abstractor coded whether the item was stated in the record as positive, stated as negative, implied as negative, uncertain, or missing (see definitions below). Five key variables were also abstracted from a randomly selected 20% subsample of charts ($n = 21$), and agreement was measured. In addition, both the chart abstractor and the principal investigator made independent, blinded judgments on overall chart legibility based on a three-point scale of illegible, average legible, and very legible.

The chart abstractor trained with the principal investigator over several training subsets of charts until they achieved a high degree of agreement. The principal investigator monitored the chart abstractor on a regular basis. Although the same research assistant who performed patient and provider interviews also abstracted the medical records, this occurred three months after the interviews to minimize abstractor recall of the interview as a source of bias.

Measurements. A variable was considered positive if it was specifically stated in the medical record as positive (e.g., ''history of CAD''). Negative variables were coded as either stated negative or implied negative. A variable was considered stated negative if it was referred to directly in the medical record (e.g., ''no history of diabetes mellitus''). A variable was considered an implied negative if related variables were recorded as positive, but not the variable of interest, and it was not noted to be positive anywhere else in the same provider note. For example, if the provider noted several positive risk factors for CAD, such as ''history of hypertension, diabetes mellitus'' and did not list as positive high cholesterol level, the variable elevated cholesterol level was considered to be an implied negative. Because we recognized that all readers might not agree with this interpretation, we measured, analyzed, and reported both implied and stated negatives separately. A variable was coded as uncertain if the area of the record referring to that variable was illegible or if conflicting results for that variable were given in different parts of the record. For example, the chest pain might be referred to as ''dull'' in one portion of the history and ''sharp'' in another. Variables that were not positive, stated or implied negative, or uncertain were coded as missing.

The time since the onset of the most recent episode of chest pain was taken at face value when it was indicated, and the duration was calculated based on the time of the interview. If the chart read, ''pain began six to eight hours ago,'' the longer time from onset was chosen. Many ED notes, instead of stating a specific time from onset, gave a general statement, such as ''pain started this morning.'' Hence, decision rules were made to define time frames for those statements. For example, ''pain on awakening'' was interpreted to occur at 6 AM. For all times, whether stated or estimated, if the duration mentioned in the record was within ±25% of that stated by the subject, the record was considered correct. Pain was also coded as steady or intermittent depending on the description in the medical record.

We analyzed subjects' positive and negative responses separately. For historical items that the subject

responded to positively, we calculated the percent of records that recorded these data as positive, falsely negative (both stated and implied), uncertain, or missing. For items that the subject responded to negatively, we performed similar measurements. If one considered the information from the prospective date the criterion standard, these outcomes would correspond to standard test characteristics.

To understand the basis for discordant data, we examined those cases where the subject interview and the chart notation disagreed and measured the percent of time the provider interview agreed and disagreed with the subject interview. The former implies a consistent history obtained by both the researcher and the provider, but the provider misreported it in the record or the research assistant incorrectly recorded it in the retrospective database; the latter implies the provider obtained a different history than the research interviewer. For data for which subjects selected a best choice out of several (e.g., location of chest pain), we simply calculated the percent distribution of responses for that variable both as found in the medical record and as obtained by the providers' histories.

Data Analysis. All data were entered into a Microsoft Access (Microsoft Corp., Redmond, WA) database and reviewed for outlying data. Corrections were made by referring back to the original data source. Descriptive statistics, including percentages, medians, interquartile ranges, and ranges, were used to summarize data. Using the subjects' prospectively acquired interview data as the criterion standard, sensitivities and specificities were calculated; for important variables, 95% CIs (exact method) are reported as well. As distinct from our model, which considered prospectively collected data the criterion standard, measurements of chance-corrected agreement in which there was not a criterion standard were measured by the weighted or unweighted κ statistic. An example of this would include the agreement between patient interviews and reinterview. Statistical analysis was performed (SAS Institute, Inc., Cary, NC).

RESULTS

Population Studied. We screened 143 patients and enrolled 107 eligible patients. The remainder was deemed unable to provide informed consent, refused consent, or had the initial decision to admit changed to discharge after their screening. All were enrolled as subjects; among them, we discovered protocol violations in three. These patients were admitted with chest pain but did not receive a ''rule-out'' MI protocol after admission. Our final study population consisted of 104 subjects. The median age of the subjects was 63 years, with a range of 30–93 years and an interquartile range of 51–77 years; 63% were men, and 86% were white. During the subsequent hospital admission, 6% of the subjects were ruled in for an acute MI, and an additional 8% turned out to have unstable angina pectoris. Among providers, 38 physicians cared for a median of two subjects each (range, 1–11; interquartile range, 1–3).

Comparison of Data Gathered Prospectively from the Subject and Retrospectively through Medical Record Review. Table 2 indicates the differences between data obtained in the ED prospectively and data obtained retrospectively. The left-hand columns list these differences when the information was positive, that is, when the subject gave an affirmative or ''yes'' response to the data item of interest. The right-hand columns list the same comparison of subject versus chart for negative or ''no'' subject responses.

Note that the third column, ''stated yes,'' corresponds approximately to the sensitivity of the chart in correctly identifying a subject who gives a positive response for the historical item of interest. For the history of CAD or of risk factors, these variables were reliably recorded. The lack of ''not stated'' responses means that all of these items or related positive responses were mentioned in the record, even if in an incorrect (''stated or implied no'') or uncertain (illegible or conflicting notations) manner. Other historical data items were recorded less well. For example, pain radiation was not reliably recorded; only 10% (neck) to 47% (left shoulder) of radiations were documented in the record. Among possible precipitating factors, only exertion was documented with any frequency.

For items for which the subject answered ''no'' (right-hand columns), the sums of the columns ''stated no'' and ''implied no'' correspond to the specificity of the chart for this item. The only group with a large number of false positives was the associated symptoms, where the provider recorded a positive history for these symptoms when the subject denied them 14%–22% of the time.

Location and Quality of Chest Pain. Among the 88 subjects (85%) who indicated that their chest pain was substernal or left chest, the medical record described this location in 44%. Although 34% of subjects described the quality of their chest pain as heaviness/pressure and another 22% described it as ache, these descriptors were indicated correctly in the medical record for 31% and 13% of these subjects, respectively.

Timing of Chest Pain. Of the 34 patients reporting steady pain, the charts reported 21% as steady and 59% as intermittent. In 20% of the charts, these data were not stated. Of the 70 patients reporting intermittent pain, the charts reported 6% as steady and 87% as intermittent.

TABLE 2. Sensitivity and Specificity of Information Found in the Medical Record Compared with the Criterion Standard of Subject Response

Information From Subject	Medical Record: When Subjects Answered ''Yes''						Medical Record: When Subjects Answered ''No''					
Variable	Prospective Answers	Stated ''Yes'' (%)	Stated ''No'' (%)	Implied ''No'' (%)	Uncertain (%)	Not Stated (%)	Prospective Answers (%)	Stated ''Yes'' (%)	Stated ''No'' (%)	Implied ''No'' (%)	Uncertain (%)	Not Stated (%)
History of CAD												
MI	32 (31)	24 (75)	0	6 (19)	2 (6)	0	68 (65)	0	16 (24)	43 (63)	8 (12)	1 (1)
CABG	16 (15)	14 (88)	0	2 (13)	0	0	87 (84)	0	13 (15)	64 (74)	9 (10)	1 (1)
History of risk factors												
Hypertension	44 (42)	34 (77)	1 (2)	8 (18)	1 (2)	0	52 (50)	5 (10)	21 (40)	25 (48)	1 (2)	0
Diabetes mellitus	14 (31)	13 (93)	1 (7)	0	0	0	89 (86)	2 (3)	24 (27)	60 (67)	2 (2)	0
High cholesterol levels	44 (42)	37 (84)	1 (2)	5 (11)	1 (2)	0	56 (54)	3 (9)	17 (30)	31 (55)	3 (5)	0
Radiation												
Left shoulder	34 (33)	16 (47)	0	8 (24)	2 (6)	7 (21)	70 (67)	5 (7)	28 (40)	12 (17)	1 (1)	24 (34)
Jaw	7 (7)	3 (43)	0	4 (57)	0	0	97 (93)	1 (1)	32 (33)	30 (31)	3 (3)	31 (32)
Neck	10 (10)	1 (10)	0	8 (80)	0	1 (10)	94 (90)	2 (2)	28 (30)	30 (32)	3 (3)	30 (33)
Back	20 (19)	7 (35)	0	7 (35)	0	6 (30)	84 (81)	1 (1)	28 (33)	27 (32)	3 (4)	25 (30)
Other	13 (13)	3 (23)	0	6 (46)	1 (8)	3 (23)	91 (88)	6 (7)	26 (29)	29 (32)	2 (2)	28 (31)
Associated symptoms												
Nausea/ vomiting	16 (15)	10 (63)	1 (6)	1 (6)	2 (13)	2 (13)	87 (84)	12 (14)	52 (60)	19 (22)	0	4 (5)
Shortness of breath	48 (46)	41 (85)	3 (6)	2 (4)	1 (2)	1 (2)	55 (53)	12 (22)	28 (51)	9 (16)	1 (2)	5 (9)
Diaphoresis	35 (34)	19 (57)	5 (14)	8 (23)	0	2 (6)	68 (65)	13 (19)	23 (34)	26 (38)	1 (1)	4 (6)
Dizziness/ weakness	36 (35)	15 (42)	2 (6)	16 (44)	1 (3)	2 (6)	67 (64)	10 (15)	9 (13)	44 (66)	0	4 (6)
Precipitating factors												
Exertion	29 (28)	10 (34)	3 (10)	0	2 (7)	8 (28)	75 (72)	4 (5)	16 (21)	5 (7)	1 (1)	46 (61)
Stress	14 (13)	0	0	4 (29)	0	10 (71)	90 (87)	1 (1)	4 (4)	33 (37)	0	52 (58)
Eating	4 (4)	0	1 (25)	1 (25)	0	2 (50)	100 (96)	1 (1)	7 (7)	31 (31)	0	60 (60)
Body position	12 (12)	0	0	4 (33)	0	7 (58)	92 (88)	1 (1)	6 (7)	30 (33)	0	55 (60)

CAD, coronary artery disease; MI, myocardial infarction; CABG, coronary artery bypass graft.

Of the 70 patients with intermittent pain, 65 recalled the time the most recent episode began. As told by the subject, the median time was 4.0 hours ago with a range of 0.3 hours to seven days and an interquartile range of three to six hours. When the medical record was abstracted, this time was neither stated nor even implied for 32 subjects (49%). Among the 33 patients with this time estimate recorded, only nine (27%) of charts agreed to within 25% of the subject-reported time.

Subject Reliability and Validity. The steps in data degradation are shown in Figure 1. Chronologically, it begins with the concept of subject reliability and validity. Among the 12 questions asked, subjects were uncertain of their response between 0 and 8% of the time. Of the subsample of 20 subjects who were interviewed twice, agreement over the 12 questions (variables) in the questionnaire ranged from 70% (quality of chest pain with seven possible responses) to 100% (e.g., histories of MI, coronary artery bypass graft, and diabetes mellitus). This corresponded to a κ statistic of 0.56 (range, 0.18–0.94) to 1.00 (range, 1.00 to 1.00). A κ statistic value of 0.56 is considered acceptable and 1.00 is considered excellent.

Clinician Reliability and Validity. These issues can best by addressed by measuring the reasons for discordance between what the subject told the prospective researcher and what appeared in the medical record. Among the 110 responses for which the subject said ''yes'' but the medical record stated or implied ''no,'' the clinician indicated that they had elicited a ''yes'' response from the subject in 28% (95% CI = 19% to 37%), indicating a mistranscription or incorrect recall of the information they received. For 71% (95% CI = 63% to 80%) of these disagreements, the clinicians indicated that they elicited a ''no'' response from the subject. In this situation, we cannot tell if the subject gave an unreliable response or if the clinician created a bias in hearing the subject's history.

Among the 83 responses for which the subject said ''no'' but the medical record stated or implied ''yes,'' the clinician indicated that they had elicited a ''no'' response from the subject in 48% (95% CI = 38% to 59%), indicating a mistranscription or incorrect recall of the information they received. For 47% (95% CI = 36% to 58%) of these disagreements, the clinicians indicated that they elicited a ''yes'' response from the subject.

Among 88 subjects with pain located in the left chest or substernal area, providers confirmed they were told of this location in 84 subjects (95%). However, the retrospective data set found this information accurately recorded in 39 (44%) of subjects, inaccurately recorded in 13 (15%), and not recorded in 36 (41%). Among the 58 subjects who described their chest pain as heaviness, pressure, or a dull ache, providers stated they were told of this description in 37 (64%). In the retrospective data set, this information was recorded correctly for 21 subjects (36%), recorded incorrectly for 11 (19%), missing for 22 (38%), and uncertain for four (7%). Hence, clinicians seem to recall a history consistent with what a patient tells a prospective researcher more accurately than they record it.

Besides these components of clinician reliability, their legibility also affects the ability of the medical record abstractor to gather accurate information. Of the 104 ED records, three (3%) were rated as illegible, 40 (38%) were rated as average legible, and 61 (59%) were rated as very legible by the chart abstractor. The weighted κ on the legibility was 0.90, suggesting excellent agreement beyond that attributable to chance.

Researcher Reliability and Validity. Like other sources of bias, it is difficult to measure bias on the part of the medical record abstractor. However, we could measure interobserver reproducibility. For the five items (history of CAD, risk factor of hypertension, location of chest pain, quality of pain, and exertion as a precipitating factor) that were coded by two investigators, the κ statistic over the 21 chart abstractions compared was 0.94 (range, 0.82–1.00) for quality of pain and 1.0 for the remaining items, suggesting excellent agreement. We did not measure lack of reproducibility because of erroneous data entry.

DISCUSSION

Loss of Information. We describe a model to describe information flow for both prospective and retrospective studies. Our principal finding was that, for data collected retrospectively through medical record review compared with data collected prospectively directly from the subject, there is a loss or degradation of information. For some items, such as the history of risk factors for CAD, the amount of data degradation is relatively small; the sensitivity of the medical record to reflect a positive history of risk factors for CAD was quite high, ranging from 77% to 93%. However, for other positive historical items, such as precipitating factors, this information was recorded in the record much less consistently. For items to which the subject responded negatively, the medical record reliably indicated these as negative for risk factors for CAD but not for precipitating factors. For many items, the loss of information was more than 50%.

Each of the three principal participants in this information flow (the subject, the clinician, and the researcher) interacts uniquely with each other, the medical record, or the retrospective data form. Thus, there are four sequential interfaces where information is transferred: that of the subject–clinician, the clinician–medical record, the medical record–researcher, and the researcher–retrospective data form. At each of these interfaces, information may be lost through data degradation. Lack of reproducibility or precision can usually be measured, whereas bias and censoring typically cannot.[6,20]

Although the discussion of the accuracy and relative value of prospectively versus retrospectively acquired data has been present in the medical literature for more than a quarter century, few studies have examined the relationship between what patients tell interviewers in real time and the availability of that information in a medical record.[1,21] Kothari et al. noted that onset of stroke symptoms was documented in medical records in only 79% of stroke patients; in more than 40% of these patients, only a general, not an exact, time was noted.[22]

Evenson et al. examined the delay from symptom onset to ED presentation in patients presenting with stroke symptoms, comparing information obtained from interviews with that recorded in the medical record.[23] In examining that subset of patients with onset of symptoms more than six hours before ED presentation, this information was recorded within a one-hour margin of error in only 30%. Using a method similar to our protocol, both the studies by Kothari et al. and Evenson et al. used decision rules to quantify approximate rather than exact times. The relatively poor capture of timing variables they describe is consistent with our results.

More recently, studies attempting to determine how well medical records reflect quality-of-care measures indicate that they document elements of care delivered to actor–patients very poorly.[24,25] All elements of the clinical encounter, including the history, were captured significantly less well in the medical record compared with a standardized-patient checklist.

Subject Reproducibility and Validity. In our study, up to 8% of subjects could not recall historical items and up to 30% gave different answers upon repeat questioning one hour after the initial interview. Hence, a corollary to the finding that information is

lost through successive steps is that even prospectively acquired information is imperfect. Because the patient may give a different response upon repeat questioning or acknowledge that they do not recall certain items, perfect information represents an ideal goal rather than a reality. We consider lack of perfect information even in a prospective study an important finding. The data degradation at this stage of the information flow would be the same for both prospective and retrospective studies.

The fact that patients as subjects poorly recall their medical history when verified by other sources is well established in the epidemiology literature.[26] Paganini-Hill et al. compared data obtained by postal surveys or interviews with that found in medical records and discovered a high level of agreement for some historical variables, such as medications prescribed or history of cancer, but poor agreement for others, such as a history of MI.[27,28] Subjects both underreported and overreported historical data. In two unrelated studies comparing disease history by self-report versus documented in the medical record, only 60%–75% of self-reports of a history of MI were confirmed in the medical record.[29,30] Kee et al. assessed the reliability of reported family history of MI as recalled by recent MI survivors and found only moderate agreement with medical records.[31] In these studies, the medical record data were obtained from entries into the medical record at several time points, not a single clinical encounter, lending it credibility as the criterion standard. Fonseca et al. compared prospectively with retrospectively acquired data and learned that, at interview, subjects recalled only 20% of injuries documented in a daily diary.[32]

The reason why subjects cannot recall their prior health information accurately is unclear. Coughlin, after reviewing the literature, concluded that a distinction should be drawn between recall that was biased and that which was nonreproducible, although this is admittedly difficult. He added that the extent of inaccurate recall is related to characteristics of the exposure of interest and of the respondents.[33] In some studies, agreement between self-reported health data and formal records seemed to be dependent on the type of historical item questioned, such as prior surgery versus pharmaceutical data.[34] Other studies have indicated that the strength of agreement between self-report and documented health data tended to be greater for male subjects than female subjects, for white subjects than African American subjects, and for subjects from referral hospitals than community hospitals. No consistent patterns were apparent by age.[35]

In addition to the concept of poor reliability, the possibility of bias on the part of the subject exists. One commonly invoked explanation for inability of subjects to accurately recall their medical history is recall bias, usually invoked in case–control studies. Kip et al., however, showed that it could play a role even in prospective cohort studies.[36] Another potential source of subject bias is the social desirability bias, described by many social scientists.[9,37] This tendency for some subjects to feel the need to give socially desirable responses to questions about health behavior is so well accepted that indices have been created to measure it.[9,38] This bias, for example, has been offered as one explanation for the underreporting of cigarette smoking noted in many health surveys.[37]

Clinician Reliability and Validity. Our data show that both illegibility and contradictory data within a single ED note occur frequently. Because information is not transmitted reliably due to these factors, they are most easily seen as ingredients of nonreproducibility rather than bias. In addition, however, our study indicated that clinicians reported false-positive results and false-negative results, and omitted including important data in their notes. Errors occurred both because clinicians obtained different histories from those obtained prospectively by researchers and because they obtained the same history but recalled or recorded it differently. For example, with respect to the location and quality of the chest pain, clinicians knew the correct data but frequently reported it incorrectly in the medical record. Although poor reproducibility may be at issue, the time between the physicians obtaining the history and recording it was brief. This raises the possibility of bias in either what the clinicians think they heard or what they chose to enter into the medical record.

Although the concept of bias toward a socially acceptable response has been invoked for subjects, providers may be influenced by similar considerations, that is, reluctance to indicate sensitive information in the medical record. Hollander et al. described a protocol similar to ours in which they asked subjects with cocaine-related chest pain what information they had given to clinicians and then tracked that information flow to the medical record.[14] Among 25 patients who recalled being questioned about cocaine use in the ED, only 44% had this information documented in their medical record.

Researcher Reliability and Validity. The final participant in this information flow is the researcher who reviews the medical record and records data into a retrospective data form. Like the subject and the clinician, the role of the researcher in transmitting accurate information may also be compromised by nonreproducibility or bias. In attempting to quantify errors in retrospective data gathering, Horwitz and Yu reviewed data extracted from the medical records of 102 patients by three trained abstractors on two separate occasions.[5] In general, they discovered high rates of intraextractor and interextractor agreement. They also categorized sources of disagreement into six

categories, among them conflicting data reported in a medical record and information not noted. We have identified these as sources of disagreement in our study as well.

In a study that examined the quality of retrospectively derived data, Gilbert et al. reviewed 244 research reports based on medical record review and published in the emergency medicine literature; specifically, they measured flaws that might affect accuracy.[4] Whereas inclusion criteria and variable definitions were commonly described in most reports, abstractor training, monitoring, and blinding as well as interrater reliability were infrequently mentioned. Recently, Worster et al. repeated a study of the deficiencies in medical record reviews in emergency medicine journals over the past ten years and concluded that, although abstraction forms and monitoring of abstractors were improved, the management of missing or conflicting data was not.[39] To address these issues, Schwartz and Panacek have published a description of the limitations of retrospective data and suggestions on how to improve its quality.[3,40] These investigators argue for, among other things, abstractor training, monitoring, and blinding; careful case selection; and strict definition of variables. They also offer concrete suggestions as to how to operationalize these recommendations.

A second implication of our model and findings is that historical data items represent different data types, such as nominal or continuous variables; the data type may play a role in how reproducibly information is transmitted. Moreover, these properties of data are unrelated to the importance of the variable in question. Hence, the accuracy with which data are recorded in the medical record may be related not just to the perceived importance of that data element by the clinician, but by the type of data that the historical item represents. Responses to most variables are either yes or no, the best choice among several, or a point on a continuous scale. The greater the number of potential responses to an historical variable, the greater the probability of mismatch between the subject's history and the information found in the research data form. The history of CAD variables was recorded with high reliability, but the quality of the chest pain when ''ache'' or ''dull'' was captured by only 13%–31% of the medical records and the timing of intermittent chest pain was captured in <25%. Because the timing of chest pain may be a relatively important variable in the chest pain story, one might expect it to be recorded with greater precision than potentially less important variables such as the presence or absence of hypercholesterolemia. However, the timing of the onset of chest pain represents a continuous variable, and the number of incorrect possibilities (even if an approximately correct answer within 25% of the true one is considered correct) is very high. Compare this with the counterexample of questions with simple yes/no responses, such as a history of cardiac risk factors or prior CAD. Here the universe of incorrect responses is limited to one. A historical item of intermediate statistical complexity between these two examples is the potential responses to the quality of chest pain, for which the true response given by the subject may be misinterpreted or mistranscribed into several possibilities. ''Ache,'' for example, may be heard by the clinician as ''fullness'' and recorded as ''pressure.''

In the model we describe, for historical items that are true, the medical record may reflect a true positive or a false negative; the opposite is true for negative items. In addition, information may be either directly stated or implied in the medical record. Most positive (true) historical items are expected to be transcribed into the medical record. These positive items are almost always stated, rarely implied. However, it is generally not the expectation that all possible negative components of the history are listed as separate negatives. Hence, the researcher using medical record abstraction needs to decide if failure to mention an item should be recorded as a negative or as missing data. A middle approach utilizes the concept that we describe as an implied negative; if diabetes mellitus and smoking were noted in the record as risk factors for CAD but no others, then family history, hypertension, and hypercholesterolemia were implicitly negative because the clinician bothered to record positive risk factors but did not list these. We have reported these implied negatives separately from stated negatives so that the reader can assess our data with or without this assumption.

Suggestions to Enhance Retrospective Data Collection. Because of the issues of cost, practicality, and rare events, among others, retrospective studies will continue to be performed. What then would help to obtain accurate data retrospectively?

Given the loss of information that occurs from each of the three participants and their interfaces, it would be helpful to limit the variability around each of them as much as possible, addressing both bias and reproducibility. This is also true for the other components of the interfaces, the medical record, and the research data form. For patients, this might mean cross-referencing information in the history of present illness with other entries in the medical record, seeking confirmation. This is particularly true for items that may be more prone to a social desirability bias, such as history of smoking.

For clinicians and for medical records, it would be best to use templates with checklists for common problems such as chest pain. Many commercially available products offer this and ensure that negative data are checked as such, rather than appearing as missing. However, hospital medical record committees may present an obstacle, and the cost of these templates needs to be addressed. Furthermore, it may

not be practical to have templates available for every complaint that is presented to an ED. Finally, attention to retrospective data entry, including double entry to reduce transcription errors, is important.

Some historical items, such as time, seemed particularly difficult to capture accurately from the medical record. Because the data type of variables at least in part determines reproducibility, we recommend that investigators pay special attention to the design of data collection tools around the more difficult variables to capture, such as time. That timing of chest pain is particularly difficult to measure retrospectively is unfortunate, because it is so necessary for the interpretation of biomarker data.

Historical items with multiple possible answers represent another challenge. Investigators should avoid making distinctions among descriptive terms that are commonly interchanged in routine clinical practice, such as ''pressure'' or ''heaviness.'' When examining the quality of chest pain, investigators would be better served to consider equivalent the more typical descriptors such as ''ache,'' ''pressure,'' and ''heaviness'' and to separate them as a group from ''stabbing'' or other more atypical descriptors.

Finally, the investigator should design the study protocol after having made a decision about the relative importance or harm of falsely negative and falsely positive information. This is simply a restatement of the standard sensitivity/specificity tradeoff. The decision of which notes to include from a medical record will affect these test characteristics. For example, considering as positive a mention of a previous MI from any one of several notes in the medical record will increase sensitivity for those patients who have experienced MIs but will decrease specificity for those who have not, because more observers have the opportunity to note a false positive. In addition, investigators should set up the database and coding instruments to reflect the fact that medical record responses to historical items may be positive, negative, uncertain, or absent and that responses may be either explicitly stated or implicit.

To address the concept of the implied negative versus missing data, it would make sense to perform sensitivity analyses around these alternative interpretations to see if the conclusion of the study would be affected.

LIMITATIONS

This study had a number of limitations. The first involves the choice of our model. A principal assumption is that data collected prospectively by trained researchers can serve as a criterion standard. Hence, we have measured the amount of data that flow into a retrospective data set in terms of test characteristics, such as sensitivity. If this prospective history is not a valid choice, then the study would be better framed as a study of agreement between retrospective and prospective data collection.

To determine the source of data degradation along the retrospective pathway, we used data obtained from an interview of the clinician as a proxy for what he or she was told by the patient. To be absolutely certain of what he or she was told, we would have needed to record the actual patient interview, risking a potential Hawthorne effect.

There were a series of other limitations. We used the patient as the unit of analysis. One could argue the physician should be the unit of analysis. Because this is largely a descriptive study with very little hypothesis testing, we did not believe it was necessary to invoke the level of complexity of using a cluster analysis. Moreover, we believe the clustering effect should be minimal because the median number of patients seen by each physician was two.

The population size was modest. However, the goal of the study was to describe the loss of information from a prospective interview to a medical record review. Hence, the point estimates in a larger study would have tighter CIs but would not likely change the conclusions. This study was conducted at a single site and an academic medical center. It is possible that other centers, particularly those where the primary ED note is written by a staff physician, might demonstrate a higher rate of recording data. However, there is no reason to believe that our institution differs significantly from other academic centers or that notes by staff would be more complete than those written by residents. Medical recording that is template driven would likely produce different results.

We did not measure interrater reliability for the prospective data collection, because the instrument used consisted of a series of closed-ended questions read to the subject verbatim and because we both trained and monitored interviewers in this process. The model we chose was chest pain because there is a well-established expectation of what constitutes an appropriate history in a patient with chest pain. We believe this model would work for any common clinical presentation. Finally, providers may have recorded historical items differently for patients who were discharged home from the ED rather than admitted to the inpatient service. However, there is no a priori reason to believe this occurs.

CONCLUSIONS

In attempting to measure the relationship of historical information obtained by prospective versus retrospective research protocols, we developed a model of information flow. Information flows through three participants—the subject, the clinician, and the researcher—at one of four interfaces defined by the participants and data-gathering tools. Information is lost in varying amounts at each of these four interfaces

through both bias and nonreproducibility. Because of subject uncertainty and inconsistency, even prospectively acquired data may be inaccurate.

Historical items may be considered as data types that affect the reproducibility with which they are recorded in medical records. Simplest variables command yes/no responses; timing is the most complex historical variable to capture.

Historical items in medical records can best be seen as positive, negative, uncertain, or missing and in addition as explicitly stated or implied. Choices among these possibilities determine the sensitivity and specificity of the medical record to reflect prospectively acquired data. Recommendations to improve retrospective data gathering are made, based on the prior work of methodologists, and are incorporated into this model.

The authors thank the house, nursing, and attending staff of our institution for cooperating with our provider interviews and data gathering; Karen Donelan, ScD, for suggestions on background literature; Shannon Lunnin for advice on manuscript preparation; and Robert L. Wears, MD, MS, for his thoughtful comments.

References

1. Sartwell PE. Retrospective studies. A review for the clinician. Ann Intern Med. 1974; 81:381–6.
2. Singer AJ, Homan CS, Stark MJ, Werblud MC, Thode HC Jr, Hollander JE. Comparison of types of research articles published in emergency medicine and non-emergency medicine journals. Acad Emerg Med. 1997; 4:1153–8.
3. Schwartz RJ, Panacek EA. Basics of research (part 7): archival data research. Air Med J. 1996; 15:119–24.
4. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: where are the methods? Ann Emerg Med. 1996; 27:305–8.
5. Horwitz RI, Yu EC. Assessing the reliability of epidemiologic data obtained from medical records. J Chronic Dis. 1984; 37: 825–31.
6. Sackett DL. Bias in analytic research. J Chronic Dis. 1979; 32: 51–63.
7. Karras DJ. Statistical methodology: II. Reliability and validity assessment in study design, part B. Acad Emerg Med. 1997; 4: 144–7.
8. Feinstein AR. Clinical Epidemiology: The Architecture of Clinical Research, ed 1. Philadelphia, PA: WB Saunders, 1985.
9. Nunnally JC. Psychometric Theory. Volume 2. New York, NY: McGraw-Hill, Inc, 1978.
10. Bosetti C, Tavani A, Negri E, Trichopoulos D, La Vecchia C. Reliability of data on medical conditions, menstrual and reproductive history provided by hospital controls. J Clin Epidemiol. 2001; 54:902–6.
11. Limkakeng A Jr, Gibler WB, Pollack C, et al. Combination of Goldman risk and initial cardiac troponin I for emergency department chest pain patient risk stratification. Acad Emerg Med. 2001; 8:696–702.
12. Paul SD, O'Gara PT, Mahjoub ZA, et al. Geriatric patients with acute myocardial infarction: Cardiac risk factor profiles, presentation, thrombolysis, coronary interventions, and prognosis. Am Heart J. 1996; 131:710–5.
13. Walker NJ, Sites FD, Shofer FS, Hollander JE. Characteristics and outcomes of young adults who present to the emergency department with chest pain. Acad Emerg Med. 2001; 8:703–8.
14. Hollander JE, Brooks DE, Valentine SM. Assessment of cocaine use in patients with chest pain syndromes. Arch Intern Med. 1998; 158:62–6.
15. Wald D, Lamden R, Curtis M. Written documentation of the chest pain history by fourth-year medical students using a simulated emergency department patient encounter. Acad Emerg Med. 2004; 11:500–1.
16. Aday LA. Designing and Conducting Health Surveys. Volume 1. San Francisco, CA: Jossey-Bass Inc, 1989.
17. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires, ed 2. New York, NY: Oxford University Press, 1996.
18. Tourangeau R, Rips LJ, Rasinski K. The Psychology of Survey Response. Cambridge, England: Cambridge University Press, 2000.
19. Judd C, Smith ER, Kidder LH. Research Methods in Social Relations, ed 6. Austin, TX: Holt, Rinehart and Winston, Inc, 1991.
20. Rothman K, Greenland S. Modern Epidemiology, ed 2. Philadelphia: Lippincott Williams & Wilkins, 1998.
21. Cummings P, Koepsell TD, Weiss NS. Studying injuries with case-control methods in the emergency department. Ann Emerg Med. 1998; 31:99–105.
22. Kothari R, Jauch E, Broderick J, et al. Acute stroke: delays to presentation and emergency department evaluation. Ann Emerg Med. 1999; 33:3–8.
23. Evenson KR, Rosamond WD, Vallee JA, Morris DL. Concordance of stroke symptom onset time. The Second Delay in Accessing Stroke Healthcare (DASH II) Study. Ann Epidemiol. 2001; 11:202–7.
24. Luck J, Peabody JW, Dresselhaus TR, Lee M, Glassman P. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. Am J Med. 2000; 108:642–9.
25. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. JAMA. 2000; 283:1715–22.
26. Harlow SD, Linet MS. Agreement between questionnaire data and medical records. The evidence for accuracy of recall. Am J Epidemiol. 1989; 129:233–48.
27. Paganini-Hill A, Chao A. Accuracy of recall of hip fracture, heart attack, and cancer: a comparison of postal survey data and medical records. Am J Epidemiol. 1993; 138:101–6.
28. Paganini-Hill A, Ross RK. Reliability of recall of drug usage and other health-related information. Am J Epidemiol. 1982; 116:114–22.
29. Colditz GA, Martin P, Stampfer MJ, et al. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. Am J Epidemiol. 1986; 123:894–900.
30. Psaty BM, Kuller LH, Bild D, et al. Methods of assessing prevalent cardiovascular disease in the Cardiovascular Health Study. Ann Epidemiol. 1995; 5:270–7.
31. Kee F, Tiret L, Robo JY, et al. Reliability of reported family history of myocardial infarction. BMJ. 1993; 307: 1528–30.
32. Fonseca SS, Victora CG, Halpern R, Lima R, Barros FC. Comparison of two methods for assessing injuries among preschool children. Inj Prev. 2002; 8:79–82.
33. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990; 43:87–91.
34. Horwitz RI. Comparison of epidemiologic data from multiple sources. J Chronic Dis. 1986; 39:889–96.
35. Linet MS, Harlow SD, McLaughlin JK, McCaffrey LD. A comparison of interview data and medical records for previous medical conditions and surgery. J Clin Epidemiol. 1989; 42: 1207–13.
36. Kip KE, Cohen F, Cole SR, et al. Recall bias in a prospective cohort study of acute time-varying exposures: example

from the herpetic eye disease study. J Clin Epidemiol. 2001; 54:482–7.
37. Brittingham A, Tourangeau R, Kay W. Reports of smoking in a national survey: data from screening and detailed interviews, and from self- and interviewer-administered questions. Ann Epidemiol. 1998; 8:393–401.
38. Okamoto K, Ohsuka K, Shiraishi T, Hukazawa E, Wakasugi S, Furuta K. Comparability of epidemiological information between self- and interviewer-administered questionnaires. J Clin Epidemiol. 2002; 55:505–11.
39. Worster A, Bledsoe RD, Cleve P, Eva K. Reassessing the methods of medical record review (mrr) studies in emergency medicine research ten years later [abstract]. Acad Emerg Med. 2004; 11:467.
40. Panacek EA. Basics of research (part 9): practical aspects of performing clinical research. Air Med J. 1997; 16:19–23.