Advanced Statistics: Understanding Medical Record Review (MRR) Studies

Andrew Worster, MD, MSc, Ted Haines, MD, MSc

Abstract

Medical record review (MRR) studies have been reported to make up 25% of all scientific studies published in emergency medical (EM) journals. However, unlike other study designs, there are no standards for reporting MRRs and very little literature on the methodology for conducting them. The purpose of this article is to provide the reader with methodological guidance regarding the strengths and weaknesses of these types of studies. **Key words:** medical records; retrospective studies; research design; evidence-based medicine. ACADEMIC EMERGENCY MEDICINE 2004; 11:187–192.

WHAT IS A MEDICAL RECORD REVIEW STUDY?

The term ''medical record review'' (MRR) refers to any study that uses prerecorded, patient-focused data as the primary source of information to answer a research question. The sources of information include: physician and nursing notes; ambulance call reports; diagnostic tests (e.g., electrocardiograms, radiographs, laboratory tests); clinic, industry, administrative, and government records; and/or computerized databases. As with any source of data for scientific study, the medical record must ''be capable of providing data that [are] both reproducible and valid.''[1] This key point is seen as the primary weakness of medical record reviews.

WHY SELECT AN MRR STUDY DESIGN?

Ideally, a study design is selected because it is the best method of answering the research question. The greatest advantage of the MRR design is that the data are already collected. Emergency health services research makes use of routinely collected data (e.g., emergency department [ED] visits and trauma databases) and chart reviews to address questions about the utilization, appropriateness, process, and outcome of care. Physicians often conduct case series studies based on reviews of charts. Most importantly, MRRs allow us to address research questions that cannot be answered in prospective trials such as: 1) the effects of harmful exposures to which people cannot be randomized; 2) the effects of potentially beneficial exposures to which people cannot be randomized; and 3) the occurrence of rare events in exposures to which people cannot be randomized.

Also, there are purposes that are impractical to address with prospective studies, such as: 1) studies of patterns of disease or behavior (e.g., ED visits) over prolonged periods; 2) quality assurance studies; 3) studies involving the sharing of cases to create large (e.g., regional trauma) databases; and 4) pilot studies to provide information for planning prospective trials.

However, using medical records will not be useful for measuring phenomena that are commonly not documented.[2] Therefore, the convenience of precollected data does not make the MRR study the most appropriate design.

HOW ARE STUDY CASES BEST SELECTED FOR MRRs?

The two primary differences between MRR and prospective studies are: 1) case selection—in the former, the cases have already occurred and the information on them is mixed with that of all the noneligible subjects; and 2) data quality—the data of MRRs were not originally recorded for research purposes and, therefore, may be lacking in quality and quantity.[1]

Prospective trials define selection criteria so as to capture potential subjects as they enter the health system (or some subsection of it) and immediately separate this cohort from all other patients. In MRR studies, the study cases need to be sorted out after they have been mixed with all of the nonstudy cases.

The presenting complaint is a frequently used case selection criterion for ED-based studies. Assuming that the presenting complaint is always recorded and

From the Department of Emergency Medicine, Hamilton Health Sciences & McMaster University (AW), and the Faculty of Health Sciences, McMaster University (TH), Hamilton, Ontario, Canada.
Received January 11, 2002; revisions received December 18, 2002, and January 23, 2003; accepted March 26, 2003.
Series editor: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, CA.
Address for correspondence and reprints: Andrew Worster, MD, MSc, Research Director, Department of Emergency Medicine, Hamilton Health Sciences & McMaster University, 237 Barton Street East, Hamilton, ON, Canada L8N 3Z5. Fax: 905-527-7051; e-mail: worster@hhsc.ca.
doi:10.1197/j.aem.2003.03.002

is done so accurately, selecting all patients based on a specific presenting complaint should be relatively straightforward.[3] However, this assumption implies that the patient, caretaker, or emergency health care worker has correctly identified the disease process for which the patient requires help, that it was recorded, and that this was done correctly as the presenting complaint. In addition, if the presenting complaint is used as the sole selection criterion to identify cases in a study, many cases of the disease of interest might be missed. Similarly, the use of the discharge diagnosis as the sole selection criterion creates a risk of missing patients who have no diagnosis listed or have more than one discharge diagnosis. This can occur because often only one diagnosis is listed on a database and there is no universal rule for which diagnosis should be listed first.

To maximize the sensitivity and effectiveness of the case selection process, one should first determine what databases are available to search (e.g., administrative, pharmacy, diagnostic imaging, laboratory) because these might provide more specific search variables, and then consider combining and cross-referencing cases. Only once a comprehensive list of *all* potentially eligible cases has been created should the inclusion criteria be applied for case selection. This will inevitably result in more work, but it will enhance the validity of the study results.

HOW TO ASSESS MRR DATA QUALITY

In prospective studies the variables are defined a priori and collected in an organized manner with quality assurance measures in place to ensure that the data are complete and accurate. This is generally not true of medical records.[1–4] In fact, each individual medical record is composed of different interpretations of different scenarios, often by different observers.[3–5] The free-text format commonly used in patients' medical charts provides additional challenges both with legibility and interpretation. These shortcomings in turn lead to a greater amount of missing and erroneous data than is generally the case in prospectively planned studies, and reduced reliability and validity of the values of the recorded variables.[3]

The quality of the data used in MRR studies is not necessarily always inferior to prospectively collected information.[1] In fact, one MRR study of patients with Hodgkin's disease demonstrated the validity of the recorded clinical information in the charts by generating survival curves similar to those of another population.[1]

Calder et al. conducted an MRR study to determine whether manually measured QT dispersion on electrocardiograms (ECGs) might be a useful diagnostic predictor of acute myocardial infarction (AMI) in patients with otherwise nondiagnostic initial ECGs.[6] In this case, whether the study was conducted prospectively or retrospectively has no impact on the data quality since ECGs provide data that are recorded essentially without subjectivity and that would be unaffected by a priori definitions or other prospective measures of data quality assurance. The retrospective nature of the data selection, however, could be argued to increase the risk of missing eligible cases; prospective data collection generally has tighter control of comprehensiveness of case identification. The study of Calder et al. shows how this risk can be minimized in that, although the authors used discharge diagnosis (AMI) for their case selection, they did so using explicit and objective screening criteria: all patients admitted to a specific observation unit combined with nondiagnostic ECGs and an elevated serum troponin level. Therefore, not only can their case selection results be easily reproduced, but their method allows for the combining of two databases (admission and laboratory), which further reduces the probability of missed cases.

Data quality in MRR studies, as we see from this example, can be high in those that use relatively objective diagnostic test data.[5] Emergency medical services (EMS) electronically recorded and uniform reporting style of times (Utstein) is another example of high-quality data that allow for sound MRR studies.[7] Using this type of data, a recent MRR study demonstrated that EMS response times less than 5 minutes were associated with improved survival.[8]

In short, there exists a spectrum of data quality such that MRR studies based on free-text data and/or subjective findings are more likely to yield less valid and reliable study results than those based on relatively objective data sources such as the examples above.

COMPUTERIZED VS. PAPER-BASED MEDICAL RECORDS

Patient-focused data can be found in both computerized and paper-based medical records. Generally speaking, computerized databases: 1) are less expensive to search since no additional manpower is required to retrieve the data from a written form; 2) are more time-efficient since data for a large population can be processed in a relatively short period of time; 3) provide more precise estimates when larger numbers can be analyzed; and 4) when available in spreadsheet format, they can quickly provide estimates of missing data.

Computerized databases can be less accurate at the level of individual patient data because of the possibility of clerical error associated with the process of transcribing the data from the chart to the database. It has also been shown that event rates can be underestimated by automated record searches unless a second record of the event exists.[9,10] However, the risk of such errors can vary with the type of event and

method of recording. The EMS automated electronic recording of times is certainly more accurate than individuals documenting times using their own watches. Also, many hospitals are now using direct computer data entry by health care workers and administrative staff. Each of these methods reduces the opportunity for transcription errors and thereby strengthens the quality of computerized data.

For individual patient data, it is recommended that where possible, the original medical chart be used as the primary source of information for the MRR or at least as a source for validation of the information obtained from the computerized database. An example of this is in an MRR study on potential thrombolysis candidates among cerebral infarction patients in which the researchers used an electronic ED database to generate a complete list of patients with the diagnosis and then confirmed the diagnosis (using established diagnostic criteria) by examining the individual medical records, laboratory results, and radiologic findings. This second search found that 117 of the original 653 patients from the ED database had diagnoses other than cerebral infarction.[11]

DATA ABSTRACTION

The onus is always on the researcher to demonstrate to the readers that the data were abstracted reliably and in an unbiased manner.[12] Data abstraction strategy is enhanced by application of a number of aspects of what is known about abstraction behavior[13]: 1) keeping the data abstractors blind to the study hypothesis decreases subjectivity in classification in relation to personal theories about the study's aims[12]; 2) the use of explicit criteria for abstracting variables results in higher inter- and intraobserver reliability because it reduces subjectivity in interpretation[1]; and 3) the accuracy of observers is increased when the individuals know that their reliability is being monitored.[14] Based on the above findings, the following data abstraction strategies are recommended in order to avoid bias and increase inter- and intraobserver reliability: 1) train the abstractors[13,15]; 2) keep the abstractors blind to the study hypothesis[12]; 3) establish unambiguous variable definitions and inclusion and exclusion criteria a priori[1,13]; 4) establish unambiguous rules regarding the management of missing or conflicting data[5]; 5) advise the abstractors at the beginning that their work will be checked for accuracy[14]; and 6) check the reliability of the abstracted data in random samples.[16]

In a comprehensive review of designing medical record abstraction forms, Banks describes in detail the many aspects that need to be considered.[17] We present some of the most important design considerations below in an abbreviated format. In creating a data abstraction form, one needs to consider the source or sources of data. If more than one source of data is to be searched for the same information, e.g., patient chart and computerized database, then a separate data abstraction form is to be used for each source. As mentioned previously, in the event of conflicting reports, one needs to have unambiguous rules regarding the primacy of data sources. To facilitate the recording of data, the questions on the form should follow the same order as the information appears in the chart.[17] Loss of data and the listing of data in the wrong category during the abstraction process can be minimized by creating appropriate categories for each variable as well as categories for missing and undetermined variables.[15] Techniques for reducing errors in recording numeric variables include providing an exact number of boxes for the number of digits (e.g., 10 boxes for a telephone number) and, in the case of single digit values in a two-box field, using a leading zero.[17]

Since duplicate data recording, i.e., first on paper and then into a computer, can also provide an additional opportunity for errors, when possible, data should be recorded directly into the computer. This also minimizes the number of omitted, illegible, or mistranscribed entries. Software programs such as Microsoft Access (Microsoft Corp., Redmond, WA) are ideal for this. A comprehensive review of this topic has been written and is a useful guide for creating abstraction forms.[17]

MISSING AND CONFLICTING DATA

Missing information can lead to nonresponse bias in the results, in that subjects with missing information may differ systematically from the others. It is an inevitable problem with retrospective studies that can range from partial information in charts to complete absence of entire charts. It has been recommended that if a given piece of information is missing from 10% or more of the cases, then that particular information should not be used.[5] There is no empiric evidence to support this and the proportion of missing data that can be accepted is dependent on numerous factors, including the study question, the type of variable, and the impact on the results. If the MRR is to be conducted on a computerized database with a spreadsheet format, one can immediately determine how much information is missing for any given variable and make appropriate allowances in the design or sample size.

Missing values are typically managed in one of two ways: 1) case deletion; and 2) imputation. Case deletion is the most commonly used method and involves simply omitting observations with missing values from the analysis. The limitations with this method are that it: 1) may introduce bias into the analysis (unless subjects with missing data are few and do not differ with respect to the outcome measures); and 2) reduces the sample size. To compensate

for missing information, one MRR study on six-month survival of patients according to their triage levels analyzed the results in two different ways.[18] In the first analysis, all patients with unknown vital status were assumed to have survived, and in the second, all those at lowest risk of death within the six-month period were assumed to have died. In this way, the researchers were able to compensate for the missing data and show that nonresponse bias did not affect their results. Similarly, if data on other predictive variables (e.g., gender, age) are available for subjects with missing outcome data, comparison can be made to determine whether they appear to differ systematically.

The second method of managing missing data, multiple imputation, is a data-driven alternative method that adjusts variances and covariances and allows for valid parameter estimates and confidence intervals.[19] Although this method uses all of the available data in the analysis, it is typically limited to very large, computerized databases.

Conflicting data, such as two or more different versions of the same event in the database, is another common issue for MRR researchers. The method of resolution of such differences, e.g., by consensus of abstractors or by accepting the first recorded observation, should be established a priori in the study protocol. Similar to the way in which published prospective clinical trials report subjects who fail to complete treatment, MRR studies should report in the results section of the publication the number of data conflicts and outcome(s) of their resolution.

SAMPLE SIZE

Once a specific research question is formulated and the MRR is selected as the best or most practical method of answering the question, a sample size should be calculated. Sample size determination is a mathematical process to decide how many subjects are needed in order to make a reasonably sound judgment about a hypothesis.[20] Sample size calculations must be performed in the planning stages of a study and are a requirement in the methods section of most peer-reviewed medical journals. How the sample size is calculated depends on the statistical tests used in the analyses. Generally, results are reported with confidence intervals (CIs) around the summary measure. Therefore, the sample size should be based on the desired CI width (usually 95%). These aspects of MRR studies are no different than with any other quantitative study except for one additional calculation: as part of the results in the MRR the intra- and/or interrater reliability of the data abstractors is also reported. Therefore, a sample size calculation for this analysis might also be performed.

The formulas for sample size calculations are found in most health research statistics books and automated methods of computing them can be found at a number of Web sites by searching the term ''sample size calculation'' (e.g., http://www.stat.uiowa.edu/~rlenth/Power/index.html). There are no published recommendations for what proportion of the abstracted data should be randomly checked for accuracy of abstraction; however, 10% is a commonly used amount, and more is better.

SAMPLING METHOD

Sampling refers to the method by which study cases or records are selected from the target population or database. Again, an advantage of MRRs is that data over a lengthy time period can be accessed all at once. A common method of sampling is to select all of the consecutive cases within a given time frame. This is a type of convenience sampling. It is an acceptable approach provided the period is long enough to include seasonal variations or other changes over time that are relevant to the research question.[21] For nonconsecutive sampling, the best method of selection is probability sampling. This provides an equal opportunity for each eligible case to be selected without bias. It is best achieved by using a random number generator (as in the MRR study on six-month survival of patients according to their triage level) or table to identify the records for selection.[18] Probability selection can be implemented across (i.e., stratified by) time periods or seasons or other categories relevant to the research question.

Two other sampling methods that should be noted are incidental sampling and systematic sampling. Incidental sampling involves selection of the most easily accessible cases from the target population. Incidental sampling is difficult to justify as being an unbiased selection method or providing a representative sample of the target population.[22] It is useful mainly in allowing an initial characterization of the types of cases. Systematic sampling involves the selection of every ''nth'' case from the target population.[22] It is considered by many to be a ''quasi'' or ''pseudo'' form of random sampling because the selection of cases is not truly random. Occasionally, when there is periodicity in the manner in which the records have been assembled, systematic sampling can lead to nonrepresentative selection. As probability sampling is not difficult to carry out, systematic sampling holds little advantage.

DETERMINING AND REPORTING RELIABILITY

Reliability is a commonly misused and misunderstood term. Rarely can MRR researchers comment on the reliability of the original data. This is another inherent problem with retrospective studies. However, researchers can measure and report the degree to which the results obtained from data abstraction

by one observation were reproduced on subsequent observations of the same record: that is, what, if any, differences were found when the data abstraction was repeated either by the same abstractor (intraobserver reliability) or by a different abstractor (interobserver reliability).[13,14,16,23] In their study of published emergency medicine MRRs, Gilbert et al. reported that only 5% of 244 studies mentioned the interrater reliability and only 0.2% tested the chance-corrected interrater agreement.[12]

Reporting the interrater reliability as a concordance rate or percent agreement between the observers is misleading when used alone because this indicates only whether the two observers had similar findings on similar numbers of records. It doesn't indicate how much of that agreement could have occurred by chance. This is the advantage of Cohen's kappa (κ) as a measure of interobserver agreement.[24] Kappa is reported as a value from –1 (perfect disagreement) to 1 (perfect agreement).[24] It is interpreted as the extent of agreement achieved compared with the total amount of agreement possible beyond chance agreement. The formula for this statistic is:

$$\kappa = [\text{observed agreement}\ (\%) - \text{expected agreement}\ (\%)]/[100\% - \text{expected agreement}\ (\%)]$$

When using the kappa statistic, it is generally recommended that researchers strive to achieve a minimum level of interobserver agreement of 60% beyond chance agreement, i.e., a kappa value of 0.6 or greater.

Although the kappa statistic is the most commonly used measure of interobserver reliability in MRR studies, it will not be suitable for all measurements of agreement. Since the kappa statistic is derived from the standard error (SE, or variance), it cannot be calculated for cases in which there is perfect agreement; hence the percent agreement between observations is acceptable.[25] Also, the kappa statistic limits the assessment of reliability to "agree or disagree" and does not measure partial agreement. For such more complex, multivariate analyses, tests such as the weighted kappa are more suitable.[25] It is beyond the scope of this paper to discuss the different statistical tests for correlation analysis and agreement, but consideration needs to be given to factors such as whether the agreement is between observers or observations, the type of variable, i.e., continuous, ordinal, or categorical, and the distribution of the variable. For these reasons, the advice of a statistician is often necessary.

CONFIDENTIALITY

Last but not least is the issue of patient confidentiality. Many medical journals require a statement of approval from the local ethics review board in the methods section of the article. Whether or not ethics boards make this requirement, safeguards should be implemented to isolate personal identifiers from the research data set and to prevent recognition of individuals in research reports. The Health Insurance Portability and Accountability Act (HIPAA) of 1996 clearly states the confidentiality rights of health insurance consumers. In essence, it states that nonidentifiable health care information should be used unless the individual has consented to the disclosure. Usually this is not a concerning issue with MRR studies as the abstracted data typically limit or exclude the case-identifying information. However, when applying for access to such information databases, the researchers are required to demonstrate that the information used will preclude any personal identifiers.

CONCLUSIONS

Medical record review studies represent a large proportion of the emergency medicine literature and recommendations for conducting and reporting them have been published.[12] However, unlike randomized controlled trials and systematic reviews, there are no clear or comprehensive standards for authors or editors to refer to in the reporting or evaluation of these studies.[26,27] The diverse nature of MRRs makes unlikely the development of consensus-based standards that would pertain to all types.[15] An understanding of the limitations and methodological issues can, however, contribute to improving the overall quality of these types of studies.

The authors thank Drs. Roger J. Lewis and Craig Newgard for their valuable input into the manuscript.

References

1. Boyd NF, Pater JL, Ginsburg AD, Myers RE. Observer variation in the classification of the information from medical records. J Chron Dis. 1979; 32:327–32.
2. Stange KC, Zyzanski SJ, Smith TF, et al. How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patient visits. Med Care. 1998; 36:851–67.
3. Burnum JF. The misinformation era: The fall of the medical record. Ann Intern Med. 1989; 110:482–4.
4. Feinstein AR, Pritchett JA, Schmiff CR. The epidemiology of cancer therapy: III. The management of imperfect data. Arch Intern Med. 1969; 123:448–61.
5. Wu L, Ashton CM. Chart review. A need for reappraisal. Eval Health Prof. 1997; 20:146–63.
6. Calder KK, Tomogin C, Mallon WK, Genna T, Bretsky P, Henderson SO. Manual measurement of QT dispersion in patients with acute myocardial infarction and nondiagnostic electrocardiograms. Acad Emerg Med. 2002; 9:851–4.
7. Cummins RO, Chamberlain DA, Abramson NS, et al. Recommended guidelines for uniform reporting of data from out-of-hospital cardiac arrest: the Utstein style. Task Force of the American Heart Association, the European Resuscitation Council, the Heart and Stroke Foundation of Canada, and the

Australian Resuscitation Council. Ann Emerg Med. 1991; 20:861–74.
8. Blackwell TH, Kaufman JS. Response time effectiveness: comparison of response time and survival in an urban emergency medical services system. Acad Emerg Med. 2002; 9:288–95.
9. Worster A, Haines T. Does replacing intravenous pyelography with non-contrast helical computed tomography result in benefit for patients with suspected acute urolithiasis in a community setting? Can Assoc Radiol J. 2002; 53:144–8.
10. Dresser MVB, Feingold L, Rosenkranz SL, Coltin KL. Clinical quality measurement comparing chart review and automated methodologies. Med Care. 1997; 35:539–52.
11. Yamagouchi K, Hori S, Nogawa S, Tanahashi N, Fukuuchi Y, Aikaawa N. Thrombolysis candidates for the treatment of stroke at an emergency department in Japan. Acad Emerg Med. 2002; 9:754–8.
12. Gilbert EH, Lowenstein SR, Kozoil-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: where are the methods? Ann Emerg Med. 1996; 27:305–8.
13. Horowitz RI, Yu EC. Assessing the reliability of epidemiologic data obtained from medical records. J Chron Dis. 1984; 37:825–31.
14. Reid JB. Reliability assessment of observation data: a possible methodological problem. Child Dev. 1970; 41:1143–50.
15. Allison JJ, Wall TC, Spettel CM, et al. The art and science of chart review. Jt Comm J Qual Improv. 2000; 26:115–36.
16. Beard CM, Yunginer J, Reed CE, O'Connell EJ, Silverstein MD. Interobserver variability in medical record review. An epidemiological study of asthma. J Clin Epidemiol. 1992; 45:1013–20.
17. Banks N. Designing medical record abstraction forms. Int J Qual Health Care. 1998; 10:163–7.
18. Wuerz R. Emergency Severity Index triage category is associated with six-month survival. Acad Emerg Med. 2002; 8:61–4.
19. Schafer JL. Analysis of Incomplete Multivariate Data, Monographs on Statistics and Applied Probability 72. Boca Raton, FL: Chapman & Hall/CDC, 1997.
20. Last JM. A Dictionary of Epidemiology, 3 ed. New York, NY: Oxford University Press, 1995.
21. Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing Clinical Research, 2 ed. Philadelphia, PA: Lippincott, Williams & Wilkins, 2001, p 336.
22. Polgar S, Thomas SA. Introduction into Research in the Health Sciences, 3 ed. New York, NY: Churchill Livingstone, 1995.
23. Labelle J, Swaine BR. Reliability associated with the abstraction of data from medical records for inclusion in an information system for persons with traumatic brain injury. Brain Inj. 2002; 16:713–27.
24. Cohen JA. A coefficient of agreement for normal scales. Educ Psychol Measure. 1960; 20:37–46.
25. Cohen JA. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70:213–20.
26. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA. 1996; 276:637–9.
27. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. Lancet. 1999; 354:1896–900.