Looking Through the Retrospectoscope: Reducing Bias in Emergency Medicine Chart Review Studies

Amy H. Kaji, MD, PhD*; David Schriger, MD, MPH; Steven Green, MD

*Corresponding Author. E-mail: akaji@emedharbor.edu.

INTRODUCTION

Chart review studies that use prerecorded data as the primary information source to answer a research question account for approximately 25% of all scientific studies published in peer-reviewed emergency medicine journals and 53% of emergency medical services journals. The popularity of the chart review study design may be partly ascribed to the fact that the data are already collected, thereby eliminating the onerous task of prospective data collection. Moreover, a chart review study allows the investigator to access data that may be lacking. For example, a researcher studying the relative unimportance, sheer oversight, or diagnostic mindset of the message often differs radically from the original. Misinterpretation of listeners seeking meaning, the distortion of the true effect estimate, and the potential for systematic error are all possible outcomes of chart review studies.

The purpose of this article is 3-fold: (1) to provide a model that identifies the numerous processes in chart review studies that can introduce bias; (2) to outline the steps an investigator may take when planning a chart review study to mitigate distortion and bias; and (3) to describe reporting techniques that optimize transparency so readers can anticipate the biases and the limitations of the study. The “retrospectoscope” in Figure 1, which we modeled after the epidemiologic telescope of Maclure and Schneeweiss, depicts the following 10 potential layers of bias, detailed below.

LAYER 1: CHART REVIEW APPROPRIATE FOR THE RESEARCH QUESTION

The chart review must be an appropriate method of data collection to answer the proposed research question. This means that the available charts must be representative of the patient population of interest, and they must include documentation of the pertinent study items. Documentation in charts is recorded for multiple reasons: billing, administrative recording, legal issues, and as a record of what transpired in the actual medical care delivered. This last assumption is key to chart review studies and often not well supported. Because data in the charts were not recorded for research purposes, comprehensiveness and quality may be lacking. For example, a researcher studying the
association between meat consumption and appendicitis could not use a chart review because the quantity of meat consumed is not documented as part of a routine medical history. The dietary history will be inaccurate if a validated and reliable means of asking the question had not been used.

Once chart review is deemed to be a valid method of data collection for the proposed research question, the availability of the data should be described. If possible, a sample size calculation should be performed and detailed. Furthermore, the investigator should not only consider whether there are an adequate number of cases but whether there are a sufficient number of cases with the outcome of interest. Additionally, if statistical modeling is going to be used, a sufficient number of outcomes per predictor variable of interest would need to be ensured. As noted above, however, the risk of random error in a chart review study design is far less than the risk of systematic bias from misinterpretation or inaccurate recording of clinical information. Thus, although a sample size calculation can help ensure that the resulting numeric study outcomes have the desired precision (ie, sufficiently narrow confidence intervals), precision does not ensure validity (ie, the absence of bias).

Unfortunately, a review of the quality of 117 chart review studies in the international emergency medicine literature revealed that only 10% reported a sample size calculation.

Solution: The first step is to determine whether the research question may be answered by using chart review as a data collection method. Establish whether the necessary data are available. Then, using a sample size calculation, determine whether there are sufficient charts to ensure adequate precision. For transparency of reporting, the study methods should describe the original purpose of the collected data.

Layer 2: Transparency of Investigator Bias

It is also important to understand and identify any investigator biases and potential conflicts of interest, be they financial or philosophical. The investigator may have unwittingly formulated a research question or a data collection instrument that is inherently skewed to favor proving his or her predicted hypothesis.

Solution: Before embarking on a study, the investigators should declare any conflicts of interest, seek institutional review board approval, and develop (and ideally pilot test) a data collection form. Each of these items should be addressed in the
Potential Pitfalls and Checkbox of Solutions

Chart review as a data collection method may not be appropriate for the proposed research question.
- Establish whether necessary data are available in the charts and whether there are sufficient charts to ensure adequate precision.
- Provide a sample size calculation.

Investigators may have a financial or philosophical conflict of interest and thereby unwittingly create a data collection instrument that is skewed to favor proving a predicted hypothesis.
- Declare any conflicts of interest before embarking on a study.
- Provide evidence of institutional review board approval.
- Submit the data collection form, as well as coding rules and definitions, as an online appendix.

The available charts may not compose a sample that is representative of the patient population of interest.
- Ensure that all eligible charts have an equal probability of selection.
- Determine and describe inclusion and exclusion criteria before data collection and analysis.
- Provide a flow diagram delineating how the study population was derived from the source population.

There may be multiple conflicting data entries in the chart, or there may be inconsistent coding of data.
- A priori, define the predictor and outcome variables to be collected.
- Develop a formal coding manual for abstractors and investigators.
- Submit the coding manual as an online appendix.

Data collection that is not systematic may lead to misclassification bias.
- Pilot test the data collection instrument.
- Use a standardized data collection instrument with clear criteria about how to handle ambiguously recorded data.

Large amounts of missing data may result in incorrect conclusions.
- Consider performing a sensitivity analysis, using methods such as multiple imputation, to encompass different assumptions to account for the missing data.

The abstractor may introduce conscious or subconscious bias.
- Select and train abstractors who are blinded to the study objectives and hypothesis.
- Describe in the article how abstractor blinding was maintained.

Interpreting chart entries and data coding and classification requires training, and non–medically trained abstractors may not know where in the chart to find information or may have difficulty resolving internal discrepancies in the medical record.
- Select data collectors with sufficient training to accurately cull the data.
- Provide uniformity of training, especially in multicenter studies.
- Describe the chart reviewers’ medical educational background, the type of training provided, and whether there were periodic refresher training sessions.

For studies involving a prolonged data collection phase, there may be a decrement in the accuracy of recording or a change in coding practices.
- Monitor and plan periodic meetings with abstractors to resolve disputes or review coding rules.

Two different abstractors may draw different conclusions from a review of the same chart.
- Perform and describe formal interrater reliability assessments, as well as the underlying agreement data.

The acceptable level of agreement between abstractors varies with the circumstances and the variable in question.
- Provide justification for the criteria for considering each variable reliable.

Few items may be checked for reliability.
- Describe the interrater reliability of each item, using an online supplement if necessary. If selected items are studied, they should be items most important to the study’s hypothesis.

There is no evidence-based standard proportion of abstracted data that should be evaluated for reliability.
- For variables that have frequent positive and negative values, an interrater reliability assessment on a 10% random sample may be adequate. This should be described in the methods section. If, however, the results demonstrate very few positives or very few negatives, describe the reliability of an enriched sample that includes an equal proportion of positive and negative outcomes.

Figure 2. Potential pitfalls in using chart review as a data collection method and checkbox of solutions.

article. Additionally, the data collection forms and coding definitions should be submitted as an appendix.

LAYER 3: STUDY AND TARGET POPULATION

A common pitfall of chart review studies is that the base population may not compose a sample representative of the patient population of interest. Studies that fail to use all or a random sample of available charts may lack internal validity. Furthermore, even if sampled charts are representative of all charts, external validity will be compromised if charts were taken from a setting with an atypical population or practice style.

Solution: Make sure that the study settings are typical of the patient population of interest, and, whenever possible, ensure
that all eligible charts are in the pool and have an equal probability of selection. The use of multiple methods of chart identification (eg, chief complaint and International Classification of Diseases codes) may help capture all eligible charts for inclusion, thereby diminishing selection bias. Inclusion and exclusion criteria should be determined before data collection and analysis. In the article, the investigator should explicitly state how the study participants were selected, as well as the inclusion and exclusion criteria for the study population. When possible, investigators should compare included charts to the target population in regard to important clinical characteristics. A flow diagram delineating how the study population was derived from the source population should be provided (see the Strengthening the Reporting of Observational Studies in Epidemiology statement at http://www.strobe-statement.org).

**LAYER 4: VARIABLES TO BE COLLECTED**

The next potential area of bias occurs in the data collection phase. There may be multiple conflicting entries. The triage nurse and resident physician may document the presence of a soft abdomen, whereas the attending physician or consultant who may have access to imaging results obtained later in the emergency department course may document the presence of tenderness or a mass. There may also be inconsistent coding of data into categories. Note that there are many opportunities for misclassification if there is any miscommunication or misunderstanding between the provider and the patient. For example, the patient may have problems with health literacy and misinterpret the question, or the provider may record what seems to fit the overall picture rather than what was actually said. There is also the possibility that the abstractor may misinterpret what is written. According to a criterion standard of reports from actor-patients, one study reported that 30% of care that was not recorded and that 19% of care that was recorded was not provided. Thus, the sensitivity and specificity of the medical record is low because there may be errors and idiosyncrasies in the reading, interpreting, coding, and transcribing of the data.

Solution: The variables to be collected from the chart, as well as how these variables are defined, should be determined a priori and documented in a coding guide for abstractors. When the methods are reported, the coding rules for each abstracted element should be provided. Additionally, when possible, the coding guide should be provided as an appendix.

**LAYER 5: SYSTEMATIC DATA COLLECTION**

Data collection that is not systematic may lead to misclassification bias.

Solution: One means to improve objective data collection is to use a standardized data collection instrument that has been pilot tested and organized and ordered in a manner similar to that in which the information may be found in the actual chart. Because each manipulation of data provides an additional opportunity for errors, when possible, data should be recorded directly into a computer program that has real-time error checking. This minimizes the number of omitted, illegible, or incorrectly transcribed entries. If, however, paper records are going to be used, then a standard case record form should be created and retained. The authors should explicitly describe the coding options. For example, if the variable of interest is the presence of rebound tenderness on abdominal examination, there are 4 potential results from a chart review: recorded and present, recorded and absent, recorded and equivocal, and no mention whatsoever. In this case, the authors should describe whether missing data were classified as such. In the article, the authors should also report whether the data collection instrument was pilot tested. If it was not, this should be discussed as a limitation. Additionally, the data collection instrument should be provided as an appendix.

**LAYER 6: MISSING AND CONFLICTING DATA**

Large amounts of missing or conflicting data may result in incorrect conclusions. Missing data can be thought of as a form of selection bias, and the degree to which selection bias can compromise validity varies from variable to variable and is determined by the context and subject matter. For this reason, there is no way to define an acceptable proportion of missing data.

Solution: When determining which variables to study, the investigators should determine the proportion of data that are missing and whether missing data threaten validity. Consideration may be given to performing a sensitivity analysis to encompass different assumptions to account for the missing data. For example, if a key variable is often missing, one could analyze the data under varying assumptions, such as the variable is missing at random; when missing, the value is always “not present”; when missing, the variable is always “present”; or some other plausible circumstance. A sensitivity analysis allows readers to understand the potential extremes in the effect estimates. The problem of missing data may be worsened if the missingness systematically differs with respect to the predictor or outcome status. Although missing and conflicting data may be dealt with by case deletion (analyzing only those cases with complete data), such an approach leads to a smaller sample size, diminished power, and potential bias, depending on associations between missingness and the states of key variables. An alternative technique is to replace missing values with reasonable replacements. The methodology for doing so is beyond the scope of this article. However, readers should understand that multiple imputation techniques based on the premise that data are missing at random may produce spurious results because this assumption is unlikely to be correct in most chart review studies. Multiple imputation and Monte Carlo simulation using Bayesian methods are other alternatives, and a sensitivity analysis comparing complete case and multiply imputed analyses will help the reader understand the effect of the missing data. Regardless of whether simple or complex analyses are performed, we recommend complete transparency in reporting the
proportion of missing data, as well as how the missing data were handled in the analysis.

**LAYER 7: ABSTRACTOR BIAS**

As would be expected, one of the greatest potential areas for bias involves the data abstractor. If abstractors are not blinded to the study objectives and hypothesis, they may be biased when assigning values for variables. Although this may not be an issue for some variables (eg, patient's sex), for other variables there may be multiple contradictory entries in the chart (eg, one physician writes that there is rebound tenderness, a second writes “no rebound,” and a third writes nothing). If the abstractors have knowledge of what the desirable value is, they may record that value, consciously or subconsciously overlooking contradictory evidence. The abstractor may also be motivated to diligently look through the chart, seeking the status of one variable, while taking the most cursory search for another. Additionally, if there are groups of patients who are being compared, abstractors should also be blinded to group assignments. Although the abstractors would ideally be blinded with respect to the aforementioned issues, only 4% of a recent series of chart review studies adhered to this methodological principle. Low adherence to the blinding principle has been attributed to the fact that the study investigators are often also the data abstractors.

Solution: The investigators should ideally select and train abstractors who are blinded to the study hypothesis. When this is not possible, an alternative is to assign different abstractors to abstract different sets of variables. For example, if one abstractor is responsible for abstracting outcomes and another is responsible for abstracting independent variables, this may decrease the possibility of bias. When reporting the methods in a chart review, the authors should explicitly state whether the chart abstractors were blinded to the study objectives and hypothesis and how such blinding was maintained. The absence of blinding should be discussed as a limitation.

**LAYER 8: ABSTRACTOR TRAINING**

Interpreting chart entries, as well as entering and coding data, requires training. Although most chart reviewers are medically trained professionals (eg, physicians, nurses, medical students), many National Hospital Ambulatory Medical Care Survey (NHAMCS) data collectors have only a high school diploma and no medical background. Non–medically trained abstractors may fail to recognize medical jargon or misinterpret test results, which can result in erroneous entries. They may also not know where in the chart to find specific information or when it is necessary to search in multiple chart locations. Non–medically trained abstractors may also have difficulty resolving internal discrepancies in the medical record. Uniformity of training is especially valuable in multicenter studies. Unfortunately, less than 20% of chart review studies describe abstractor training. Electronic records are becoming more common and, although the standardization in such records may make certain abstraction tasks easier, the idiosyncrasies of such records (eg, some information may be tucked away in obscure locations) can complicate abstraction efforts unless abstractors have intimate knowledge of the system.

Solution: Investigators should use data collectors with sufficient training to accurately cull the data. There should be uniformity of training, especially in multicenter studies, and, for large studies, refresher training should be planned. The methods section should describe the chart reviewers’ medical educational background, in addition to the type of training the abstractors underwent, whether it was standardized training, and whether periodic refresher training was provided. If there was no standardized training, this should be discussed as a limitation.

**LAYER 9: ABSTRACTOR MONITORING**

For studies involving a prolonged data collection phase, data collection forms should periodically be compared with the actual medical record charts because, over time, there may be a decrement in the accuracy of recording or a change in coding practices.

Solution: Meetings with the abstractors may be useful to resolve disputes or review coding rules and should therefore be planned. Whereas one study advocated using 3 points during the chart audit phase for quality monitoring, there is no evidence-based standard for the appropriate frequency of abstractor performance monitoring. In a recent study, only 9% of a series of chart review studies documented such monitoring. The methods should describe whether the abstractors were monitored, how frequently they were monitored, and who performed the monitoring.

**LAYER 10: ABSTRACTOR INTRARATER RELIABILITY**

Ideally, 2 abstractors would independently analyze each chart so that differences could be identified and resolved. This, however, is time consuming and costly, and hence is seldom done. The alternative is to establish the intrarater reliability of abstractions so that the results of a single abstraction of each chart can be trusted.

Intrarater reliability assessments are particularly important when there are groups of different abstractors, such as in a multicenter study. Without a comparison of abstractions by different groups, it is impossible to know whether differences among sites are due to differences in the data or differences in the abstractions.

Solution: Formal intrarater reliability assessment should be performed, with the methodology described in the methods and the results presented in the results. For studies extending over a prolonged time, such intrarater reliability should be periodically reassessed.

**LAYER 10A: AGREEMENT OR RELIABILITY**

Both raw agreement and chance-corrected agreement (reliability) should be reported. Unfortunately, chart review
publications often fail to report interrater reliability. Raw agreement can be misleading when reported alone because it does not indicate how much of that agreement could have occurred by chance. In contrast, Cohen’s $\kappa$, which is a measure of chance-corrected interobserver agreement for categorical data, is reported as a value from −1 (perfect disagreement) to 0 (chance agreement) to 1 (perfect agreement). There are many limitations to using $\kappa$ to characterize interrater reliability. A weighted $\kappa$ should be used for multiple categories because a simple $\kappa$ describes only binary categories. For continuous variables, an intraclass correlation coefficient should be reported. Each of these measures is a condensation of the actual agreement matrix, and a table displaying these underlying values is the best way to convey agreement, either in the article or in an online supplement.

Solution: The investigators should report both reliability and agreement statistics, as well as provide the underlying agreement data.

**LAYER 10B: WHAT IS GOOD ENOUGH AGREEMENT OR RELIABILITY?**

Landis and Koch provided oft-quoted criteria for interpreting $\kappa$: however, the acceptable level for $\kappa$ varies with the circumstances and the variable in question. Therefore, the use of their criteria is discouraged. For example, if the outcome variable of interest is death, perhaps no less than a $\kappa$ of 1.0 should be achieved. Authors should discuss why a certain level of agreement or interrater reliability is presumed acceptable, according to the nature of the variable in question.

Solution: Authors should justify their criteria for considering each variable reliable.

**LAYER 10C: WHAT ITEMS SHOULD BE CHECKED FOR RELIABILITY?**

It is not enough to simply state that an interrater reliability assessment was performed. If the investigators used a 30-item data collection instrument, the $\kappa$ for the assessment of reported age may be 1.0, but it would be more important to understand how often 2 raters agreed on the most important explanatory variables, the most important confounders, and the most important outcome measures.

Solution: Ideally, the interrater reliability of each item should be reported, using an online supplement if necessary. When selected items are studied, they should be the items most important to the study’s hypothesis.

**LAYER 10D: WHAT PROPORTION OF THE DATA SHOULD BE CHECKED FOR RELIABILITY?**

There is no evidence-based standard proportion of abstracted data that should be evaluated for reliability. Many studies will sample 10% of charts; however, this may or may not be an adequate sample, depending on the circumstances. For example, for a variable that is usually answered one way (eg, 95% of the time the answer is no), a 10% sample may be insufficient because there will be few or no yes answers in the reliability sample, and readers will have no idea whether abstractors are able to record yes answers accurately. In this case, the interrater assessment should ideally be taken from an enriched sample (eg, a sample that is composed of a sufficient number of both positive and negative assessments).

Solution: For variables that have frequent positive and negative values, interrater reliability assessment on a 10% random sample may be adequate. If, however, there are very few positives or very few negatives, consider instead analyzing an enriched sample that includes equal proportions of positive and negative outcomes. The methods and results sections should denote whether a random subset or an enriched sample was used for interrater reliability assessment.

The principles outlined in the 10 layers above also apply to large public and proprietary databases, such as the NHAMCS, which are derived from chart reviews. The authors who use these databases for publications, however, are not directly involved in abstracting or entering the data for collection and analysis. Thus, the authors of the article may not have access to the original data collection instrument and may not know the exact proportion of missing data. Furthermore, with these large databases, it is often unknown who the abstractors are and what type of medical background they have, how the abstractors were trained, whether they were monitored, and whether there was any interrater reliability assessment performed. To achieve optimal transparency, authors who use these large databases should provide information about the abstractors and reliability assessments according to what is known about the database from published works or from communication from the creators of the database.

**LIMITATIONS**

This article provides suggestions on the conduct and reporting of studies in which chart review is used as a data collection method. Our major recommendation on reporting is to be transparent and report exactly what was done and what was found for each issue discussed in articles. This recommendation comes directly from principles associated with the scientific method, which emphasizes the importance of describing a study in sufficient detail to permit replication. Our recommendations on the conduct of chart review studies, although created with full knowledge of the limited evidence that is available, are essentially opinion. Thus, there is no guarantee that adherence to any or all of the recommendations will produce a less biased article.

The ultimate effect of electronic medical records on retrospective research remains unclear. New biases may be introduced through the use of boilerplates, items copied and pasted, default tick boxes, and delays in time stamps relative to actual care.

The outline above and the attached checklist are thus a summary of best practices from the literature. We believe that authors of chart review studies should report all of these items and, when they do not, should explain why in the limitations section.
The authors acknowledge the Annals of Emergency Medicine panel of methodology and statistics editors for their critique of this article and their many valuable comments.

Supervising editor: Michael L. Callaham, MD

Author affiliations: From the Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, CA (Kaji); the Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA (Kaji, Schriger); the Los Angeles Biomedical Research Institute at Harbor-UCLA, Torrance, CA (Kaji); and the Department of Emergency Medicine, Loma Linda University Medical Center, Loma Linda, CA (Green).

Funding and support: By Annals policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships exist.

Publication dates: Received for publication February 28, 2014. Revisions received March 18, 2014, and March 25, 2014. Accepted for publication March 26, 2014.

REFERENCES