Using Pre-existing Databases for Prehospital and Disaster Research

Samuel J. Stratton, MD, MPH

Research using data derived from pre-existing databases is a popular method for health research. Examples of pre-existing databases include hospital patient databases, prehospital services databases, government health system databases, and databases maintained by nongovernment organizations and the World Health Organization (Geneva, Switzerland). It is a developing standard for medical and health data to be maintained in databases. As a result, database research is more frequent and robust. A search of PubMed, the medical literature search engine for the US National Library of Medicine, shows greater than 56,000 references that present or discuss research derived from databases.[1] Databases are effective for compiling large numbers of cases which makes them popular for prehospital research. Detailed country health databases maintained by the World Health Organization have application to disaster health research.

Using data recorded in a database provides a researcher ready access to study information without the need for prospective data collection or the time required for retrospective review of health records. Database research allows for collection of large numbers of cases which is useful when studying rare occurrences or diseases. For both prehospital and disaster research, database research may be an effective way to study events and exposures that cannot be prospectively studied or randomly applied to a population because of ethical reasons.[2] Database research is particularly appropriate for prehospital and disaster research because an emergency event or disaster cannot be controlled or easily predicted making prospective studies problematic.

Use of databases for research has many of the same challenges as those for chart-review-based research.[2-4] As with classic chart-review-based research, database study is prone to systematic error that can lead to selection bias, marginal error, and interpretation bias (Table 1). Well-designed databases require that a number of steps be taken to assure data are accurate and reliable. In addition, a researcher that uses a database must be disciplined in use and interpretation of prerecorded data that are derived for a study. This discussion explores the primary actions that are required for reliable database study and evaluation.

Studies that utilize pre-existing databases are, by definition, retrospective. Retrospective studies have inherent risks of systematic error. Selection bias is the most recognized risk for retrospective analysis. Selection bias is error based on the inappropriate inclusion or exclusion of subjects into a study (database). Examples of factors that can lead to selection bias include variation in individual availability for study, lack of well-defined and applied inclusion and exclusion criteria, exclusion due to missing information, or subtle issues such as exclusion due to language spoken. For database research, missing data may lead to selection bias because those excluded from analysis due to missing data may represent important individuals in a proposed study population. Databases may also lack information for all variables that may affect a study outcome (confounding variables), such as data concerning chronic disease states that could lead to poor outcomes for persons evacuated from disaster zones. Information bias is another risk for retrospective database studies, because data entered into a database may be imprecise or poorly defined as well as potentially invalid due to data entry error or sloppy data entry. Another form of error found in prehospital research is Berkson bias, which refers to only those responded to by Emergency Medical Services being included in a prehospital study and not an entire population. Prehospital stroke research illustrates Berkson bias in that prehospital studies of acute stroke reflect only a portion of a total stroke population because many stroke victims are taken to hospitals by family members and not by the Emergency Medical Services system.

Missing data within a study database are an important research problem. Missing data allow for selection bias by allowing preference for study subjects with complete data. Often, subjects with missing data are excluded from a study. Excluding otherwise eligible study subjects results in an unknown influence to data analysis and introduces a margin of error for final study results. To address missing data, researchers commonly provide the demographic profiles for those subjects with complete data and those with missing data with the hope that there is no difference between the two. Statistical imputation of datasets can be performed to adjust for missing data elements. Imputation is the process of replacing missing data with substituted values by statistical means. Efforts to avoid and adjust for missing data help researchers, but missing data must be realized as representing unknown individual elements in a database and therefore contributes to outcome measure error.

More difficult to address than missing data is invalid data entry. Validation of data entered into a database is essential if data analysis is to be considered accurate. Invalid data entry can occur from several sources, including error due to variation for data elements without absolute definition, data entry error, and miscoding error. Database elements must have a predetermined definition of value or meaning to allow for interpretation. For example, the vague term "soft tissue injury" can include anything from bruising to complex lacerations. Precise definition of database terms must be established before a database is formed and is usually maintained as a data dictionary. Data entry personnel may also make mechanical errors in entering data, either placing data in a wrong cell of the database or accidentally omitting or substituting data. Similarly, data entry personnel may misunderstand data definitions or raw data notations and miscode data as they are entered. Invalid data are difficult to manage once entered into a database and are best limited by assuring the ability of data entry personnel to accurately record data into the database. Training data entry personnel in the exact interpretation of raw data and the precise definitions used for the

https://doi.org/10.1017/S1049023X15000011 Published online by Cambridge University Press

Challenges with Conducting Pre-existing Database Research or Evaluation
1. Retrospective Research Methodology
2. Missing Data Elements
3. Data without Absolute Definition
4. Data Entry Error
5. Miscoding Error
6. Misunderstanding Error (patient or institutional)
7. Study Design Errors
8. Interpretation Error by the Researcher
9. Abstraction Error
10. Time Synchronization Error (for time studies)

Table 1. Challenges with Conducting Pre-existing Database Research or Evaluation

Essential Information to Provide in a Research Report When Using a Pre-existing Database
1. Source or Reference Citation for Database
2. Source or Reference Citation for Database Data Dictionary
3. Source(s) for Raw Data Included in Database
4. Inclusion and Exclusion Criteria for Data Entered Into Database
5. Published or Reported Accuracy and Reliability of Database
6. How Missing Data is Managed
7. Training or Known Accuracy of Database Data Entry Personnel
8. Training for Study Data Abstractors
9. How Data Abstraction Accuracy was Measured
10. Ethics Committee (Institutional Review Board) Review Outcome

Table 2. Essential Information to Provide in a Research Report When Using a Pre-existing Database

database is a preferred method for assuring reliable data entry. Other means to assure valid data are determining and minimizing rate of data error by comparing a sample of data entered by one person with that entered by another data entry person using the same raw data. For highest validity, using the technique of double-data entry is preferred. Double-data entry is the entering of the same raw data by two persons into two versions of a database and then electronically comparing the two database versions for inconsistencies. When analyzing database data, a researcher must be wary of data that are inconsistent with the majority of data and, if possible, refer back to raw data sources for possible confirmation of suspicious data elements.

Raw data for databases are often from multiple sources. This can lead to errors of misunderstanding how to classify data for input into a database. General terms within a database such as "injury," "cardiac ischemia," and "infection" are vague and often not precise enough to draw conclusions regarding outcomes. As noted in the paragraphs above, exact definitions of terms used in a database are essential and precision in definitions preferred. Often, standardized classification systems are used to help decrease misunderstanding. Frequently used standard classification systems include the Cerebral Performance Classification (CPC) for cardiac resuscitation research and the International Classification of Diseases (ICD) system.[5-6] Whether using standardized data entry systems or definitions coded in a data dictionary, a valid database must be developed such that misunderstanding of data terms is minimized.

Pre-existing databases are often developed for quality monitoring, financial analysis and billing, and as administrative tools for management, and not with research as a primary objective. For these reasons, a study designed with intent to use a database must be appropriate for the data available. Data-dredging or designing a study around impressions of the data in a database (starting with data and then formulating a study question) is not appropriate. A researcher must be sure when asking a study question and using a database that the data available will potentially answer the question in a significant manner. Frequently known confounders to an outcome that are not recorded in a database makes testing a hypothesis or study question with available data meaningless. When a database is used for research, study design should allow for appropriate use of the data available.

Researcher data interpretation error and abstraction error refer to two potential problems that occur with using a database for study. Researchers may misinterpret data or over interpret data that are recorded in a database. This often occurs when a researcher is not familiar with the data dictionary and takes recorded data for face value as interpreted by the researcher. As already noted, the data dictionary provides precise definition of data elements and should be considered the only valid interpretation of meaning for data elements. A similar error can occur with data abstraction from a database when data are entered into a study without reference to the precise meaning for the data as described in the data dictionary.

A final error that is of importance to prehospital and disaster research is time entry error. Prehospital research often focuses on time intervals. Yet, time intervals in databases may lack precision when different clocks (or watches) are used to record different time elements. Unless all clocks are synchronized to a standard time reference, time data may be imprecise and prone to significant error. Similarly, during disasters, it is not uncommon to measure events in terms of days from a primary event. How a day is measured can vary with preference that one point in the twenty-four hour clock be identified as the start of a day; for example, a day is measured as 0700 hours to the following 0659 hours. While estimates of dates and time of an event may be helpful for overview of a time-sensitive measure, estimates are prone to error that may not be acceptable for disaster research. This problem has been noted in research regarding limb amputations after the 2010 Haiti Earthquake in which estimates from time of injury to time necessary for amputation were often estimated and vague.

A few final points should be taken regarding research based upon pre-existing databases. First, all research should undergo Ethics Committee review for protection of human rights.

https://doi.org/10.1017/S1049023X15000011 Published online by Cambridge University Press

Studies using databases that contain any form of personal identifiers must be reviewed by a study Ethics Committee. Second, a researcher should always provide information regarding the validity of a database in the Methods Section of a research report. This information should include that described in Table 2, including information about the training and reliability of database data entry personnel and study data abstractors as well as any published measures of reliability of the database.

References

1. Pubmed. US Library of Medicine National Institutes of Health Web site. http://www.ncbi.nlm.nih.gov/pmc/. Accessed December 26, 2014.
2. Kaji AH, Schriger D, Green S. Looking through the retroscope: reducing bias in emergency medicine chart review studies. *Ann Emerg Med*. 2014;64(3):292-298.
3. Worster A, Bledsoe RD, Cleve P, Fernandes CM, Upadhye S, Eva K. Reassessing the methods of medical record review studies in emergency medicine research. *Ann Emerg* Med. 2005;45(4):448-451.
4. Gilbert EH, Lowenstein SR, Kiziol-McClain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: where are the methods? *Ann Emerg Med*. 1996;27(3):305-308.
5. Stiell IG, Nesbitt LP, Nichol G, et al. OPALS Study Group. Comparison of the Cerebral Performance Category score and the Health Utilities Index for survivors of cardiac arrest. *Ann Emerg Med*. 2009;53(2):241-248.
6. International Classification of Diseases (ICD). World Health Organization Web site. http://www.who.int/classifications/icd/en/. Accessed December 30, 2014.

doi:10.1017/S1049023X15000011