VALIDATION TECHNIQUES FOR MEDICAL DATA

Bob Brown
Medical Association of Georgia
938 Peachtree St., N.E.
Atlanta, Georgia 30309

Bob Harbort
Emory University Hospital
Atlanta, Georgia

Kriss Lattimer
Whittaker-Medicus
Houston, Texas

Susan Peake
Grady Memorial Hospital
Atlanta, Georgia

ABSTRACT

A medical data base is only as good as the data it contains, and the process of collecting medical information and converting it to machine-readable form is fraught with opportunities for error. An analysis of the regularities in medical data permits the development of validity and reasonableness checks which can be automated. We examine the regularities to be found in common elements of medical data and suggest checks based on these regularities. Particular attention is given to constructing checks which lend themselves to automation.

We draw on experiences from the medical records department of a 1,000 bed teaching hospital and three multi-hospital quality-of-care review projects.

INTRODUCTION

A medical data base is only as good as the data it contains, and the process of collecting medical information and converting it to machine-readable form is fraught with opportunities for error. Yet all too often the process of "editing" medical data consists of little more than verifying that each item is of the required type. Far more extensive checks are possible because of the regularities which exist in medical data. A system designer can take advantage of these regularities to construct validity and reasonableness checks. The application of such checks permits the detection of many types of errors, offering the opportunity to improve data quality.

There are several reasons why data quality is important. When medical data are used in a clinical setting, the quality of the data has a direct effect on delivery of care to the patient. The medical record is not only a record of the changing status of a patient, it is also an instrument of communication among the health care practitioners concerned with the patient's diagnosis and course of treatment. Inaccurate or incomplete data in the medical record are a serious threat to the patient.

A corollary to practicing good medical care is monitoring its delivery to make sure good practice standards are maintained. The medical record and the data in it are the main instruments of quality assurance reviews. Inaccurate or incomplete data cannot document adequately the quality of the care actually delivered. Accrediting agencies require assurance of the quality of care. Good data quality is a necessary component of both internal and external reviews of quality of care. Reimbursing agencies can demand accountability for care delivered, and this is likely to increase with growing concern over health care costs. There may also be issues of legal liability.

Finally, there is a growing recognition that good corporate business practices are applicable to operating hospitals. This requires uniformity of data among corporate substructures so that corporate decision making and planning may be carried out effectively, and so that actions within and among departments will not be inconsistent.

Data quality control is a prerequisite to quality assurance and utilization review procedures. To be of use, medical data must be accurate, timely, complete, and retrievable. These characteristics of medical data affect its use in both active (clinical) and retrospective (statistical) data bases.

We consider three general areas where errors may arise: in charting, in abstracting and coding, and in data conversion. In each case we consider the ways in which errors arise, how they may be characterized and measured, what might be done to detect them, and how they affect the use of data.

We consider two general methods of checking data: validity checks and reasonableness checks. Validity checking is the determination that the value of a data item is a member of the set of valid values for that item. Reasonableness checking is the determination that the value of a data item falls within reasonable bounds in the context in which the item occurs.
The Nature of Errors

Sources of error in medical data begin with the medical record. The three ways in which charting errors may occur are charting erroneous data, not charting data that should be charted, and incorrect or inadequate record linkage. Errors in primary charting also cause problems with the retrospective use of medical records. Aggregate analysis or audit of charts requires retrieval of comparable cases and this is made much more difficult by missing data items or sections of the chart. Inadequate linkage data may make information about entire episodes of care inaccessible.

The opportunity for error in the abstracting of medical data is enormous. The abstracting process is a filter for the purpose of data reduction, and it has the effect of changing the detail with which we perceive the data. Individual medical records are the written representation of unique human experiences; abstracting reduces our perception of detail to a point at which we are willing to classify sets of records as "the same." This process is not simply one of restating the data, but results in loss of information.

Aside from the usual errors of omission and commission, the major problem which may arise in abstracting is that of inconsistency. If different abstracting procedures are followed for two different records, the abstracts will be incommensurable; moreover, there will be no way to determine that an abstracting error has occurred. This is especially true of slight variations in technique, which may be both extremely harmful to data correctness and almost impossible to detect.

Coding has many of the same problems as abstracting, but it is even more sensitive to errors of omission. Coding represents the delivery of medical care—a continuous process—as a set of discrete elements. Errors of omission at this point cause loss of information, which usually cannot be reconstructed by inference because of the discrete nature of coding systems. The person doing the coding is frequently not the person delivering the care, which makes it very hard for coding to reflect accurately all the subtleties of the delivery of care.

Finally, opportunities for error arise when medical data are converted to machine-readable form for computer processing. In some cases medical data are captured for machine processing automatically by instruments which generate machine-readable results. More frequently, however, data from a chart or abstract are converted by keying. The types of keying errors which can occur are errors of omission, substitution, and transposition. In the latter cases, only one or two characters in a data item are erroneous. Like abstracting and coding, keying is generally performed by someone other than the person who delivered the care. Moreover, keying is often performed from a document which is the result of an abstracting and coding process, and by a person with little or no medical knowledge.

The importance of this is that errors in medical data are likely to result in systematic errors in the delivery of medical care to patients. The medical record is an instrument of communication. It is used to pass information among those involved in a single episode of care, as well as to communicate serially between episodes of care. The harmful effects of errors in medical data are most obvious when those data are used for clinical decision-making, but it is also clear that errors in data used for retrospective analyses may lead to delivery of less than optimum patient care over an entire population.

How Errors Are Detected

The key to detecting errors is redundancy. If the same information is represented in two or more places, or in two or more ways, then a comparison may be made among these representations. The more redundancy which exists in the data, the greater are the opportunities for uncovering errors. In cases where there is a great deal of redundancy, it may even be possible to devise plans for automatic error correction.

Redundancy may be introduced deliberately. The procedure of "key verification" involves having two operators key the same data, then comparing the results. This is quite expensive, and may not be cost effective for some applications. Fortunately, there exist patterns or regularities in medical data items, and in the relationships among them, which provide a certain amount of "built in" redundancy. It is possible to construct validity and reasonableness checks based on these patterns and relationships. The designer of a data set can introduce redundancy purposely by measuring two or more related variables. This is frequently not only less expensive than the 100% redundancy of key verification, it may also be better because the redundancy is introduced at the time data are collected.

In a hospital where each patient's sex is shown in the chart as either "male" or "female" we can state categorically that any other entry is in error. This is a validity check. Given a patient's chart showing a temperature of 98.6, we really have two items of information. The first is the number recorded. The second is the implicit information that this number represents the temperature of a human being. Because we know that 98.6 is the normal temperature for a human, we can assert that this value is "likely to be correct". On the other hand, in the absence of information that the patient's body temperature was reduced artificially, we could assert that a value of 89.6 is "likely to be in error." This is a reasonableness check.
Validation Checks

We defined validity checking as "the determination that the value of a data item is a member of the set of valid values for that item." A great many items of medical data lend themselves to validity checking. In particular, any item that is represented as a code can be checked for validity. The fact that a certain entry is a member of the code does not mean that it is "right" but an item which is not a member of the code set is by definition wrong. A "sparse" code set (i.e., one in which the set of entries is small compared to the number of possible entries) offers more redundancy and greater opportunity for error detection.

In addition to formal coding schemes like diagnosis or procedure codes, many other types of information have well-defined sets of valid values. We have given gender as an example. Dates and valid years can be specified as well.

Validity checks may also be constructed using data from two or more items. For example, you can verify that a recorded admission date is not later than the recorded discharge date for the same episode of care. Virtually every date in a collection of medical data can be subjected to this type of verification. It is possible to define subsets of diagnosis and procedure codes which are applicable only for patients of a certain sex or within a certain age range. When two or more related measurements are made, it may be possible to construct validity checks based on the nature of their relationship; e.g., systolic arterial blood pressure can never be less than diastolic.

Reasonableness Checks

It may happen in medical data that a value for a data item will be "valid" in the sense discussed above, but still represent a condition which occurs so infrequently as to be almost certainly in error. A diagnosis of breast cancer in a male patient is an example. (Contrast this with a diagnosis of prostate cancer in a female patient, which we would call "not valid." ) Many laboratory tests produce results which lend themselves to this kind of editing.

Reasonableness tests can be divided into three sub-types. The first is a check for static values which are generally applicable. For example, a potassium level of 7.0 is "unreasonable" regardless of who the patient is. The second type takes into consideration the patient population. A reported age of 87 would be viewed differently if reported in the records of a sports-medicine clinic than in a study of Medicare patients. In the third type, trends in an individual patient are considered in light of the body's capacity to support such changes within the reporting interval. If a patient's daily weight were reported as 152, 153, 125, 132 we would reject this as unreasonable even though any one of those values taken out of context might be reasonable.

It is important to remember that while an exception to a validity check is unequivocally wrong, this is not the case with an exception to a reasonableness check. One may find that an exception to a reasonableness check is representative of the patient's condition, and a way must be provided to inform the validation process that the unreasonable item is indeed correct.

Finding candidates for reasonableness checks is easy; setting the limits for those checks is often less so. In many cases it will be possible to establish limits through the exercise of medical judgement. In other cases it may be more appropriate to use statistical methods to set limits. This is especially true in population-based checks.

How to Implement Validation Techniques

Implementation of these validation techniques requires a stepwise approach. The first requirement is the definition of the data set with due regard to the intended purpose of the system. The designer must specify how each item in the data set will be collected. Having specified his data collection methods, the system designer will be able to estimate the degree of accuracy and consistency with which data will be collected. For each item in the data set, consider how accurately it may be obtained, how important it is to the system's operation, and the costs of techniques for validating the item and correcting any errors. In this way, a cost/benefit ratio can be developed for each item in the data set. You may not be able to quantify this ratio for each item, but that must not preclude consideration of the costs and benefits involved.

It is at this point that data validation techniques may be designed into a system. Validity checks should be applied first, followed by reasonableness checks. In each case apply single-item checks first, then cross-field checks. Careful consideration must be given to the case of a cross-field check in which one of the items has been determined to be suspect as the result of a previous check. Also, remember that it is necessary to have a mechanism to override reasonableness checks.

Don't specify validation techniques because they are easy to automate. Construct the techniques you need and then figure out how to automate them. Difficulties in automation are likely to arise not because of the complexity of a process but because of lack of rigor in its definition. If you have been rigorous in specifying your techniques, their automation will be straightforward; otherwise you force system implementors to guess at what you want.

Once the data system is operating, you can apply statistical sampling techniques to measure the quantities which were estimated as a part of the system design. If the measured values are found to be markedly different from the estimates you can adjust the system to adapt it to the setting in which it is operating.
CONCLUSION

The quality of data in a medical information system can have a profound impact on the delivery of care, regardless of whether the data are used in a clinical setting or for retrospective purposes. To be useful, such data must be accurate, timely, complete, and retrievable. Yet in the process of collecting medical information and converting it to machine readable form opportunities for error abound. By taking advantage of the redundancies in medical data and analyzing the regularities which exist in such data it is possible to detect many of the types of errors which can occur.

We have described two major types of validation techniques: validity checking and reasonableness checking, and suggested areas in which each can be applied in medical data. We have noted three possible classes of reasonableness checking and suggested applications for each, and we have commented on the process of setting limits for each class of check.

A word of caution is in order. We have discussed ways of assuring the correctness of data. However, controlling the correctness of data does not control the appropriateness of its use. For example, assurance that a birth date is correct does not make it usable as the sole vehicle for record linkage. The system designer must give attention to the intended use of data as well as to the quality of data items to ensure overall quality in a data system.

BIBLIOGRAPHY


Erickson, Jon J. "Quality Assurance of Medical Computer Systems" Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care. IEEE. 1981.


