Measuring the Quality of Medical Records: A Method for Comparing Completeness and Correctness of Clinical Encounter Data.

Judith R. Logan, MD, MS[1], Paul N. Gorman, MD[1], Blackford Middleton, MD, MPH, MSc[2]
[1]Division of Medical Informatics and Outcomes Research, Oregon Health Sciences University
[2]Medicalogic/Medscape, Inc., Hillsboro, Oregon

This paper explores the attributes of quality in recorded clinical encounter data, examines issues in measuring these attributes, and describes a method for measuring two attributes, completeness and correctness. The method is defined in the context of computer-based records and is demonstrated in a pilot study. Videotaped physician-patient encounters and an empiric process of determining a gold standard for content are used. The methodology was found to be feasible. Problems encountered during the pilot study can be remedied.

Introduction

Computerization of the medical record does not put to rest enduring questions about the quality of the data it contains. Of the different parts of the record, it is the clinical encounter data – that record of the patient's history and physical findings and the clinician's assessment and plan of care arising out of a unique interaction with a clinician – which presents the greatest problems for recording methods. Does the record of the clinical encounter accurately reflect the content of the patient's visit? Is the information complete? Is it correct? Is it a valid indicator of the quality of care provided? Questions such as these are at the core of any discussion of quality in medical records, and ensuring the quality of the record is essential whether it is being used to care for patients, to train health professionals, to conduct research, or to manage the health system.

The quality of medical records may be defined in various ways and described in terms of a number of attributes, depending on the perspective and purposes of the user. In order to improve the quality of medical record systems, it is important to define and measure these attributes. For example, various methods of documenting the clinical encounter – pen and paper, dictation, free text typing, use of encounter forms or templates – may influence the nature and quality of the data that is recorded. Only by defining and measuring the relevant attributes of quality can we compare different medical record systems and examine the influence of various system components, on the quality of the data those systems contain.

The first goal of this work is to develop a measurable gold standard that can be used for comparison of the clinical content of various records. The second goal is to describe a study methodology which will, while comparing recording methods, isolate as the single independent variable the method of data entry, controlling wherever possible for other sources of variation, such as that due to the patient, the clinician, the clinical context, etc.

Defining a measurable standard for quality

Development of this method requires a dissection of the term "quality" as it applies to patient records into constituent attributes, a choice of the attributes of quality to be examined, and operational definitions of these attributes that permit their measurement. Quality in medical records has been described as having the attributes of legibility, accuracy, completeness, and meaning[1]. Use of computer based systems can potentially lead to improvements with respect to each of these attributes, from the improved presentation of data, to the use of constrained choices and data validation rules that reduce data entry errors, to standardization of the core data elements for a record to encourage completeness.

There is precedent in informatics research for the use of completeness and correctness (used here synonymously with the term accuracy) as measures of medical record quality. Hogan and Wagner[2] in a review of studies of record quality appeal for the uniform use of completeness and correctness in future studies. They define completeness as the proportion of observations in the gold standard that are included in the record, and correctness as the proportion of the included observations that have the correct value. Others have agreed in principle with these concepts, although differences remain in their measurements having mainly to do with the level of granularity of the elements measured.

Ultimately, the quality of a patient record must be judged by whether or not that record serves the purpose for which it was intended. While the true gold standard may be the same for all of its purposes, that the record reflects the state of the patient, the optimal number of data elements, optimal granularity of these elements, and optimal presentation in the record may vary according to the purpose.

The patient encounter must therefore be analyzed by data elements and some type of gold standard established for the presence and value of those

1067—5027/01/$5.00 © 2001 AMIA, Inc.

elements in the encounter record. We suggest that the recorded data elements and the types of errors that can occur in comparing these elements to the gold standard can be classified as follows:

n1: element present in the gold standard, present and correct in the report (a Correct Element);

n2: element present in the gold standard, present but incorrect in report (an Incorrect Element);

n3: element present in the gold standard but not present in the report (a Missing Element); or

n4: element not present in gold standard, but incorrectly present in the report (an Extra Element).

A classification of data elements and errors similar to this has been used by other authors, and is proposed here as being comprehensive and mutually exclusive. Combining the text descriptions for completeness and correctness given by Hogan and Wagner with the classification above, working definitions of completeness and correctness can be derived. It appears that completeness, the proportion of observations in the gold standard that are actually recorded in the CPR, is given by:

$$completeness = \frac{n1+n2}{n1+n2+n3}$$

while correctness, the proportion of the recorded observations that are correct, relative to the gold standard, is given by:

$$correctness = \frac{n1}{n1+n2+n4}.$$

Each data element then takes on two dimensions. Elements cannot simply match the gold standard, but must be classified first as present or not and then as correct or not. It is possible, for instance, for all elements to be present in a record (completeness = 100%) while the record is completely inaccurate (correctness = 0%).

The calculation of completeness and correctness now depend on reaching agreement about the gold standard, i.e. on defining the individual, countable, atomic units of data in the clinical encounter. Options for data elements found in the informatics literature include summary information – such as problem lists, diagnoses, or keywords – specific historical items such as medications or treatments, predefined standards for the content of examinations, or established criteria for evaluation of specific diseases. However, none of these approaches allows for evaluation of overall documentation over a wide variety of patient encounters.

In defining the units of information to be measured when looking at overall documentation, some studies start by defining sets or categories of medical data and proceed to judge record content based on these sets. For example, Romm and Putnam[3] measure the presence of data in "units" which are usual in a medical encounter, such as "respiratory unit". Zuckerman et al[4] use items found typically in medical audits. Norman et al[5] begin with examining the content of individual records and then determine the critical information and actions necessary for achievement of acceptable performance for each encounter. Pringle, Ward and Chilvers[6] include as data elements "topics" or "items" which are defined when both the physician and the patient "use at least one phrase or sentence in its discussion; or if a prescription review takes place without mentioning the underlying topic explicitly." Moran and associates[7], in addition to determining units for overall documentation, weight those units as being very, moderately, or not significant in relation to the complaints stated by the patient.

With these various approaches, the resulting granularity of the data elements varies, from the low level of detail of 5.5 topics per encounter in the Pringle study[6] to the higher level of detail of 28 to 54 items per case in the Norman study[5]. This variation illustrates a problem with inferences, abstractions, and summarizations in the study of quality in clinical records. This task is not as simple as counting recorded facts; a method must account for the interpretation of or justified inferences drawn from the facts, as well as summative statements about those facts. Some degree of inference may be justified, but other summative statements can overstate or oversimplify the case or combine elements that are more appropriately stated separately.

Because these issues are unlikely to be resolved in the near term, we propose that the gold standard for data elements for a patient encounter be determined empirically by domain experts viewing the encounter. We furthermore encourage that these elements be expressed with a high degree of detail. That is, rather than defining an ideal level of detail for elements a priori, one begins by listing the significant findings of the encounter; the gold standard data elements are then defined by this list of findings. Using this approach, an observation is significant – a finding, in the terminology of Evans and Gadd[8] – if it is felt to be so by the clinicians who might make use of the record. It has the further advantage of being able to accommodate inclusion of items at varying levels of abstraction or summarization. What is an "element"? An element is a piece of information that independently adds meaning in the record.

In abstraction of records for presence of equivalents to these determined standard elements, the term "equivalent" must be applied liberally. Precedence for this liberal interpretation can be found in other informatics work and depends significantly on clinical judgment. A summation of or inference from an element can be considered acceptable as long as it is not contrary to other elements. A useful approach is whether, from a particular record, a clinician would know that a particular element is present. If so, then the element should be counted as present.

Controlling Variability

Many sources of variation are present in the interaction between patient and clinician and in the documentation of this interaction in the medical record. To isolate the factors of a medical record system as independent variables and minimize variation due to other sources, we suggest videotaped patient encounters as the preferred patient material. This material can then be viewed by a series of subject clinicians in settings, which mimic office practice. This choice assures that every subject clinician is given identical material to record and decreases the variability of results based on anything but the recording method. Videotapes should be able to reflect the patient encounter, including gestures and nonverbal cues which may convey important information. With videotapes and verbalization of physical findings by the examining clinician, it should also be possible to accurately convey findings such that both history and physical examination portions of the encounter record can be tested.

The use of videotaped patient encounters for assessing the quality of medical evaluation and medical records is not new. Residency programs often videotape the resident staff encounters in order to review and critique their skills, and other studies have used videotapes of patient encounters to evaluate the quality of record keeping. However, with some exceptions, other works have used videotapes only once, rather than reusing them with several clinicians as suggested here.

Several other methods of presenting patient encounters could potentially be employed. Standardized patients have been successfully integrated into clinical practices under evaluation with a low rate of detection. While presumably more realistic, this method adds as a confounder some variation in content of the patient-physician interaction. This method also requires recruiting and training of standardized patients, reducing feasibility and adding expense. Viewing of the actual patient encounter by one or more observers has also been employed, where the research involves comparison of records made by the observer(s) and by the clinician who participates in the encounter. With this method, multiple observers can be used as a check on interrater reliability. While this method has the advantage of using a true-to-life clinical encounter, it is not readily scalable for use with multiple subjects, and requires either the intrusive presence of observers during the interaction or facilities such as one-way mirror exam rooms to permit unobtrusive observation. Staged encounters may be repeated for multiple observers, but this again poses significant technical difficulties.

One might question whether or not a videotape accurately enough reflects a patient encounter to use for this purpose. Videotapes have primarily been used in the past to study the process, not the content of care. It is our suggestion, however, that until a better standard can be developed, videotapes are the best option for this study design.

A Pilot Study Demonstrating the Methodology

In this pilot study, we compared encounter data recorded using two methods: dictation with subsequent transcription and Logician™ Encounter Forms. Logician™ is a computer-based outpatient record system produced by MedicaLogic, Inc. (now Medscape, Inc.) of Hillsboro, Oregon. It was chosen as the CPR for this study because it provides the ability to capture data in a fully structured fashion, as free text or as templated free text, for the ability to customize forms within the system, and because of the local availability of clinicians who routinely use it in their practices. For this study a demonstration version of Logician™ (v4.2.1) was installed and all data was recorded on a single laptop computer, which allowed for control over the forms used and for storage of the data, and did not interfere with the clinicians' production versions.

Clinical encounter data can be captured in Logician™ using three types of encounter forms: blank notes, Note Templates and Encounter Forms. Each of these allows entry of free text, whether transcribed from dictation or through keyboard entry. The Note Templates and Encounter Forms possess a structure that can serve as a prompt for data. Both also permit the use of predefined, "boilerplate" text, as well as automatic entry of certain data from prior visits, and templates can be printed for producing handwritten notes. The Encounter Forms provide the additional feature of structured data entry, with radio buttons, action buttons, check boxes, drop-down lists, flowsheet views, and single and multiline edit fields. These features allow for much flexibility in clinical data entry, with potential impacts on the quality attributes being measured.

We used two Encounter Forms in this study. The "Multiple SOAP Note" is a previously developed set of forms that contain a series of labeled multiline edit fields, but no predefined text. It is designed for general use in followup visits. The "General History" note is a series of forms adapted from preexisting forms such that it includes only a history section. It provides highly structured data entry, particularly for the Review of Systems section, as well as single and multiline text entry fields. Free text entry is available as an alternative even for elements that are included as individual fields for structured data entry. Although in practice, the clinician would have significant flexibility in choice of the Logician™ forms used, exercise of this choice would preclude meaningful study comparisons. For this pilot, therefore, we chose to narrow subjects' choices to the preselected forms, as well as the Medication, Problem, Prescription and Allergy lists.

For dictation of the encounters, a standard handheld audiorecorder was used with subsequent transcription of the tapes.

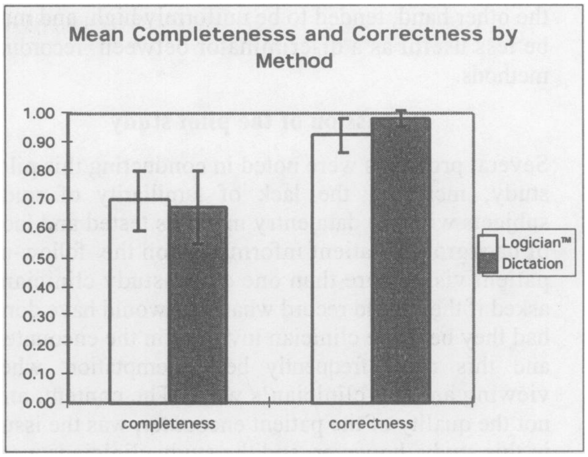
A convenience sample of eight clinicians (seven physicians and one nurse practitioner) was recruited to be the study subjects and blinded to its purpose. Five had used Logician™ regularly in their practices for periods of from 2 to 15 months. All but one had significant prior experience with Logician™, either in practice or research, and all could be considered intermediate to expert computer users. For the one clinician who had no experience with Logician™, a brief tutorial was provided prior to the study session.

The clinician subjects who used Logician™ in their practices were surveyed as to their normal use of the software. It is interesting to note that although these clinicians had access to the highly structured Encounter Forms, none of them reported using them. An average of 84% of charts created by this group were partially or totally dictated notes without use of templates, with the balance recorded by keyboard entry into blank notes.

Four videotaped patient encounters were obtained for this study; two staged encounters and two actual patient encounters, with appropriate patient consent. Two of these recordings were used only for training purposes as described below. The other two were used for data collection and consisted of one new patient visit and one follow-up patient visit for multiple medical problems.

The gold standard for data content of the patient encounters was determined by consensus of an expert panel consisting of three experienced clinicians. These clinicians individually viewed the videotaped encounters, one or more times as needed, and listed the elements which they felt should be included in the clinical record. The lists created were then pooled and returned to the panel members who were asked to view the encounters again and to confirm or deny the appropriateness of data on the pooled list. There was little disagreement after this second round and so a third round was not conducted. Each clinician was interviewed by one of the authors to clarify the elemental status of each piece of data. The encounter for the new patient visit was found to have a total of 63 elements, and the encounter for the follow-up patient to have 27 elements.

In test sessions with individual study subjects, each clinician viewed a training videotape first, then both of the test videotapes. Immediately after viewing each videotape, the clinician recorded the encounter either by dictation or into the appropriate Logician™ Encounter Form. The order of test videotape presentation and the recording method used were assigned at random, with each clinician recording one test encounter by each method. This produced eight records for each of the two patient visits, four for each using dictation and four using Logician™.



The records were transcribed or printed, then abstracted by the one of us (JL) looking for the presence and correctness of the elements which were found in the gold standard. A second abstraction by another physician who was not an author was performed on a percentage of the records as a check on the reliability of this process. The results were in agreement on 92.2% of the elements, with a kappa statistic of 0.82 overall – 0.90 for one patient and 0.65 for the other – indicating a moderate to high degree of concordance. Only the single abstraction was used in the results reported below.

Results

Mean completeness and correctness for the two methods are compared in the Figure. For completeness the mean ± SE using dictation was 0.677 ± 0.127 (95% CI, 0.55 – 0.804), while using the Encounter Form mean completeness was 0.69 ± 0.45 (95% CI, 0.592 – 0.806). For correctness, the values using dictation were 0.982 ± 0.012 (95% CI, 0.953 – 1.0) and for the CPR were 0.926 ± 0.025 (95% CI, 0.866 – 0.985). These differences were not statistically significant.

There were too few observations to perform an analysis of variance, but simple inspection reveals substantial variation in the completeness values beyond that attributable to recording methods. Between-clinician variation was significant, with mean scores ranging from 0.55-0.90. Within-clinician variation was also present, with little variation in completeness for 2 subjects and substantial variation in completeness for 3 others. Correctness values, on the other hand, tended to be uniformly high, and may be less useful as a discriminator between recording methods.

Discussion of the pilot study

Several problems were noted in conducting this pilot study, including the lack of familiarity of study subjects with the data entry methods tested and lack of background patient information on the follow-up patient visit. More than one of the study clinicians asked if they could record what they would have done had they been the clinician involved in the encounter, and this must frequently be a temptation when viewing another clinician's work. The content, and not the quality of the patient encounter, was the issue in this study, however, and the study clinicians were reminded of this. Finally, some means of accounting for correct but extraneous data in the records must be devised; in our results, extraneous data was ignored.

Conclusions

To measure the quality attributes of completeness and correctness, an overall design has been suggested here which takes into account and seeks to reduce the effect of sources of variation in the data. The method described is likely to be most useful when the clinical encounter involves a relatively clear-cut problem that matches the specialty domain and experience of the clinician. There is likely to be more disagreement about which data should be collected and recorded with poorly formulated problems or when complicating medical, social, and personal factors are at play. Similarly, there is likely to be greater variation in the data collected and recorded when the expertise or experience of the clinician are not well matched to the clinical problem, since physicians with greater knowledge and experience in a problem domain require less information to reach an appropriate diagnosis.

This method of measuring completeness and correctness is meant to permit evaluation and comparison of methods of recording the data, and is not meant as a means of assessing the content of the record itself, nor the quality of the care provided. What is claimed in this work is that completeness and correctness should be measured in a consistent manner and remain an essential part of the evaluation of quality in computer-based patient records.

The methodology described here was tested in a pilot study and found to be feasible. Problems encountered during the pilot study can be remedied.

Acknowledgments

MedicaLogic, Inc, Hillsboro, Oregon for use of Logician™ software. Dale Kraemer, PhD for advice regarding study design and analysis.

References

1. Institute of Medicine. The Computer-Based Patient Record. An Essential Technology for Health Care. Revised ed. Washington, DC: National Academy Press; 1997.

2. Hogan WR, Wagner MM. Accuracy of Data in Computer-based Patient Records. JAMIA 1997;4:342-55.

3. Romm FJ, Putnam SM. The Validity of the Medical Record. Medical Care 1981;19(3):310-5.

4. Zuckerman AE, Starfield B, Hochreiter C, Kovasznay B. Validating the Content of Pediatric Outpatient Medical Records by Means of Tape-Recording Doctor-Patient Encounters. Pediatrics 1975;56(3):407-11.

5. Norman GR, Neufeld VR, Walsh A, Woodward CA, McConvey GA. Measuring Physicians' Performances by Using Simulated Patients. Journal of Medical Education 1985;60(December):925-34.

6. Pringle M, Ward P, Chilvers C. Assessment of the Completeness and Accuracy of Computer Medical Records in Four Practices Committed to Recording Data on Computer. British Journal of General Practice 1995;45:537-41.

7. Moran MT, Wiser TH, Nanda J, Gross H. Measuring Medical Residents' Chart-Documentation Practices. Journal of Medical Education 1988;63(November):859-65.

8. Evans D, Gadd C. Managing Coherence and Context in Medical Problem-solving Discourse. In: Evans D, Patel V, editors. Cognitive Science in Medicine. Cambridge: MIT Press; 1989.