Communication of Clinically Relevant Information in Electronic Health Records: A Comparison between Structured Data and Unrestricted Physician Language

by Philip Resnik, PhD; Michael Niv, PhD; Michael Nossal, MA; Andrew Kapit; and Richard Toren

Abstract

We report on a study comparing unrestricted physician dictations with structured input in electronic health records. The results suggest a need for attention to the contrast between physicians' naturally occurring language and the information permitted by structured data entry. We suggest that technology that automatically populates structured EHR fields directly from physician dictations, rather than controlling physician input, may answer the needs of the EHR without sacrificing physicians' ability to fully communicate clinically relevant information.

Introduction

It is widely believed that electronic health record (EHR) systems will be "critical to the delivery of consistent, high quality health care."[1] Widespread adoption of EHRs will change healthcare in ways that we can only imagine at this point, ranging from more effective information sharing to biosurveillance, monitoring of interventions, and medical research based on large-scale analysis of clinical records.

In order for EHRs to be effective for many of these purposes, raw clinical data will need to be transformed into clinical information. The distinction between these concepts is familiar in knowledge management and information science within a framework known as the DIKW (data, information, knowledge, wisdom) hierarchy (e.g., see Cleveland[2]). Data can be characterized as raw or unprocessed—the pixels on a screen before the objects take shape for us, or the words that constitute a physician's dictation. Information arises from data by means of adding structure and categories in order to create units—for example, the set of physical conditions in a patient's clinical history, or the list of medications the patient is taking. Knowledge arises when we identify relationships between units—for example, identifying how a patient's prior conditions influence current problems, or how a combination of medications affects the outcome of treatment. Finally, wisdom is the ability to make good choices, informed by our knowledge and by our ability to assess the quality of our own predictions based on the knowledge we have.

Interoperability between systems that use EHRs requires shared understandings at the level of information. The structure and categories assumed by one system must correspond to the structure and categories assigned by a different system if they are to work together. As a simple example, if the category *PatientID* is conceptualized as name plus date of birth in one system and Social Security number

in a different system, then it will be difficult or impossible to achieve a shared, comprehensive view of a patient's information that takes the information from both systems into account. Similar concerns apply with respect to clinical categories. For example, a researcher might be interested in medical records labeled at the information level as involving a patient who suffered a myocardial infarction, regardless of whether, at the data level, that event had been described as *myocardial infarction*, *MI*, or *heart attack*.

In order to achieve commonality at the level of information, one proposed approach is to change the way physicians interact with health information systems: instead of entering traditional clinical dictations expressed in natural language, physicians are asked to structure and categorize at the time of input. In effect, this approach transforms the physician's process from semi-structured creation of data to structured entry of information. For example, the EHR system might require that the term *myocardial infarction* be chosen from a list, and given that entry, it might prompt the physician for additional information known to be associated with that diagnosis. We will refer to this form of information input interchangeably as structured or discrete data entry.[3]

Structured data entry has advantages. For example, it eliminates or at least drastically reduces the problem of variable terminology, by forcing the input of a standardized term or concept. This means that information technology systems developed in accord with the standard will have a much easier time interacting with each other. Structured input may also foster more complete gathering and entry of information, since a structured, systematic process, like any checklist, helps to ensure that necessary steps are not skipped; this can potentially lead to improved quality of care.[4]

At the same time, structured data entry has its disadvantages. Trachtenbarg, based on personal experience and discussion with more than 50 organizations, comments that "clicking or typing text multiple times is generally slower than dictating," also citing Waegemann and colleagues.[5, 6] He also observes that "discrete data may not catch the nuances of patient variability" and recommends "using discrete data selectively rather than trying to use it for everything."[7, 8]

Trachtenbarg's point about using discrete data selectively raises an important question. Although EHRs generally include opportunities for free-text entry, we have already seen that much of the value in electronic records comes from standardization of information, and unrestricted free-text fields work against the ability to completely standardize the information in the record. Therefore one can expect that as EHRs become more widespread, the demands for standardization may lead to pressure on physicians to forgo free text wherever possible. If physicians restrict themselves primarily to structured data entry, what happens to the "nuances of patient variability" to which Trachtenbarg refers? To state the question more generally, what information is lost when free dictation of data is replaced with structured entry of information?

To our knowledge, nobody has yet attempted to answer this particular question; it appears rather more typical for studies of EHRs to focus on costs, economic value, and effect on quality of care,[9] all of which are important but none of which bear directly on the question of what information is conveyed or not conveyed by the EHR. In this paper, therefore, we report on an initial study designed to assess the information in standard, narrowly structured EHR formats as compared with dictated free text.

Experimental Design

A set of 20 naturally dictated cardiology notes was sampled from a collection of 2,000 notes.[10] Language in each dictation was manually highlighted independently by two annotators, with text being highlighted if it conveyed information that would be captured in structured data entry according to a government-mandated EHR standard currently in use.[11, 12] Inter-annotator consistency was assessed by computing per-word agreement on the decision to highlight or not highlight for content-bearing words.[13] Word-level agreement on the notes was 91.4 percent, establishing that human annotators can agree reasonably on an interpretation of what language would and would not correspond to information included in this EHR standard's structured information. For example, in the sentence "I recommend a stress dual-isotope nuclear study in the very near future to look for possible exercise-induced myocardial ischemia," the test is included, as is "recommend," because the specification states that tests can be

categorized as events, recommendations, requests, and the like. However, the specification makes no provision for encoding hypothesized conditions for which the test would be seeking to obtain evidence, nor for encoding vague temporal specifications like "in the very near future."

An independent MD cardiologist and a cardiology expert (the "experts") each independently reviewed all 20 notes.[14] Their task focused entirely on communication of clinically relevant information; EHRs were in fact not mentioned in the instructions. For each note, each expert was asked to imagine himself as a physician assuming responsibility for the patient, and to imagine that the highlighting had been added by the previous physician, indicating what he or she believed to be clinically relevant and necessary to include in the communication. *Clinically relevant* was defined to mean "anything pertinent to the patient's current condition or future care."

Each expert identified all spans of text that he felt should have been highlighted as clinically relevant, but had not been, and assigned a 1–5 rating indicating the seriousness of each omission. The instructions defined 1 as indicating omissions of "minimal severity" and 5 as meaning that "failing to mark up the language was extremely severe, in terms of having serious consequences for the care of the patient if that clinically relevant information had not been communicated to you." The experts spent an average of about three hours on this task.

Results and Analysis

The primary question of interest here is whether, and to what extent, the naturally occurring dictations contain clinically relevant information that does not fit into discrete (non-free-text) fields specified in the EHR standard. By highlighting information that would be included according to that standard, and then asking experts to identify omitted information, we provide a way to answer this question quantitatively.

Our analysis looks at omitted text spans identified by the MD cardiologist, and we also consider spans for which both experts' independent judgments agreed. We break out three categories for omissions: a seriousness rating of three or greater (□ □),four or greater (□ □),or five (□) The percentages in Table 1 represent the proportion of 20 notes for which at least one omission occurred in that category. The next row shows the average number of such omissions per document.

For example, the second column in Table 1 shows that 100 percent of the notes were identified by the MD cardiologist as containing at least one omission rated with severity of 3 or higher, with 5.25 such omissions on average. Moreover, he found omissions with "serious consequences for the care of the patient" (severity rating equal to 5) in fully 55 percent of the notes. The first column in Table 1 uses a more conservative analysis criterion, considering only omissions that both experts identified independently, and categorizing severity using the minimum of their two ratings. (For example, an omitted span rated 4 by one expert and 3 by the other would count toward the □ □category, not □ □.) Even with that more conservative criterion, 50 percent of the notes contain at least one omission that *both* experts agree is of seriousness 4 or greater on a 5-point scale.

Now, it is possible that the EHR specification we used[15] could be extended straightforwardly to include some kinds of missing information. For example, that specification includes no field capturing negative patient reports of symptoms (e.g., "denies any chest discomfort"), but such fields could be added.

What is of greater interest, however, is information that seems difficult to include within any discrete entry framework. Analyzing the omissions flagged by our experts, we found several categories of clinically relevant information that appear to be uncontemplated in terms of discrete data entry, and for which it would seem challenging to extend current EHRs, even in principle.[16]

One such category includes nuanced or detailed elaborations of information. For example, the cardiologist found it relevant that after identifying reporting severe pain in one patient's neck and back, the dictating physician adds that she was "almost brought to tears just in getting her up on the examination table." Both experts found it relevant that a patient was "able to walk on flat levels and walk at a moderate pace for one hour without abnormal shortness of breath or chest pain."

Another category is temporal or logical context. For example, both experts found it relevant that a patient's nonsustained ventricular tachycardia (fast heart rate) occurred "during post myocardial infarction care…far removed from the time of his infarction." The cardiologist found it highly relevant, for another patient, that the dictating physician was "hesitant to recommend his FAA certification renewal" without a repeat of a previous catheterization.

A third category includes relevant information about the dictating physician's thought process. In one case, the physician recommends continuing Toprol because it "seems to be controlling [the patient's] palpitations well." In another, the dictating physician considers discomfort to be "suggestive of angina." In a third, the dictating physician expresses a belief that results of stress testing "would rule out significant major coronary artery disease, despite it being a somewhat incomplete study."

The second column in Table 2 shows the results of an analysis in which we considered only omissions only of these kinds—that is, omissions it would seem difficult to remediate in the EHR without doing violence to the entire concept of discrete data entry. This analysis is strictly more conservative than the analysis reported in Table 1, since it assumes that parts of the dictation would be included in discrete data entry even if they are not part of the current EHR specification. Yet even on this more conservative criterion, we find that, according to the cardiologist, fully 45 percent of the notes—nearly half—contain at least one omission that would seriously affect the care of the patient, rated 5 on a 5-point scale. Every note but one (95 percent) contains at least one omission that he would rate with severity 3, 4, or 5.

Finally, we adopt a still more conservative version of this analysis, restricting our attention to difficult-to-remediate omissions *and* considering only missing information that both experts agree is clinically relevant (Table 2, first column). For severity ratings of 3 or higher, there is at least one such omission in 50 percent of the notes, and over the full data set they occur on average about once per note. The experts agree that in 25 percent of the notes, there is at least one piece of information omitted whose consequences for the patient's care would merit a seriousness rating of 4 on a 5-point scale.

Structured Representations without Structured Input

In the previous section we saw that clinically significant information can be omitted if physician communication is limited to the structured information within a standard electronic health record rather than a full dictation. This presents us with something of a dilemma. On the one hand, many of the benefits of electronic health records depend on the fact that they contain standardized, structured encodings of patients' information. On the other hand, if clinicians enter structured information rather than free dictations, important information may be lost.

A possible solution to this dilemma emerges from the recognition that *structured representations* are not the same thing as *structured input*. The benefits of EHRs depend on the discrete information they contain, not on physicians entering information discretely. Within the technical discipline of natural language processing (NLP), a core interest of many researchers and developers for two decades has been *information extraction*, the automatic identification of information in unstructured text,[17] including techniques explored in clinical domains.[18] The abstraction of information for high-accuracy automated coding[19] already contains many of the elements needed in order to successfully map unrestricted dictations to standard code sets for diagnoses and procedures, as well as many other information categories identified in the EHR.

A potential alternative to restricting physicians' input, therefore, is to allow physicians to focus on the care of the patient, and express themselves clinically as they always have, using natural, unrestricted language. NLP software can automatically derive structured information that satisfies the need of many applications for consistent, discrete representations.

This approach has the advantage of preserving a key property of the data-information-knowledge-wisdom hierarchy, namely the fact that knowledge creation is a cyclical process. At any given time, the information categories we choose in our structured representations are determined by our current state of knowledge. By preserving the physicians' original language, the data, it becomes possible to return to medical records retrospectively with new knowledge in hand and to reanalyze it according to new information categories.

As an example, physicians started describing *ground-glass opacity*, a kind of hazy opacity within the lungs, with the introduction of 64-slice CT scans around 1991.[20] In principle, 16 years' worth of medical records could be mined for clinical connections related to this phenomenon. But would we have access now to all that information if those medical records had been created via structured input, rather than free dictation? It seems unlikely: the concept (as "ground glass appearance") first appeared in a standardized nomenclature (MedDRA) in 2001. If clinicians had entered information by selecting among existing information categories, this concept would not have been there to select for 10 years.

Conclusions

In this paper we have performed what is, to our knowledge, the first systematic comparison between free natural language dictations and the information codified by structured categories within an EHR. There are several limitations that should be discussed. First, our interest in natural language processing gives rise to the possibility of experimenter bias. To address this, we have erred on the side of caution whenever possible. For example, not only do we establish reasonable inter-annotator agreement for the task of extracting information relevant to structured EHRs from free dictations, but we consider information to be highlighted if *either* annotator highlighted it. Similarly, one could reasonably argue that an omission matters if either expert flagged it, but in our "both experts" analysis, an omission is only an omission if it was identified independently by *both* experts, and its severity category is based on the *minimum* of their two ratings. We go even further by including an analysis, in Table 2, that gives the EHR credit for discrete information types that *could* be added to the standard even if they are not currently present. Even under these most conservative assumptions, we find that 50 percent of the notes include at least one omission rated 3 or higher on a 5-point scale, and 25 percent contain omissions rated 4 or higher (Table 2, first column).

A second limitation of this study is its size, which is limited by the overwhelmingly manual nature of the component tasks. We find it likely that a larger and broader sample, going beyond cardiology notes, will reinforce the main point more strongly. In 20 notes we found a wide variety of clinically relevant information that was not anticipated in an EHR standard as being relevant. We predict that a larger sample will expose an even wider variety.

A third limitation of this study is the fact that we manually annotate text in dictations that would correspond to an EHR's discrete data fields, in effect simulating an EHR via its specification, rather than observing what physicians do in practice. This limitation is also one of the study's primary virtues, since it abstracts away from specific EHR implementations and permits a direct comparison of discrete information (or information that can be made discrete) with unrestricted dictation. At the same time, it raises the question of whether physicians would use free-text fields available in EHRs to augment discretely entered information, mitigating the omissions we found.

This is a sensible question to ask. Trachtenbarg's comments on the relative efficiency of dictated versus structured input[21] lead us to suspect that the latter will discourage users from troubling to enter free-text elaborations, nuances, and reasoning that would have shown up in their naturally occurring dictations.

More important, though, the existence of free-text fields is itself a concession to the need for EHRs to carry clinically relevant information that cannot be captured using discrete fields and predefined nomenclatures. Given that such a need exists, it is imperative that we evaluate the impact of structured data on inter-physician communication, considering potential information lost as well as potential value added.

In this paper we have begun by asking to what extent naturally occurring dictations contain clinically relevant information that would be lost using discrete (non-free-text) fields specified in an EHR standard. Even under quite conservative assumptions, we have found that important clinical information, detail, and nuance would fail to be captured by an EHR standard's discrete fields, with potentially serious consequences for the patient. Such omissions could potentially influence not only clinical care, but the progression from data to information to knowledge discovery in clinical research. Clearly the question merits further attention and study.

Philip Resnik, PhD, is a strategic technology advisor for CodeRyte, Inc., in Bethesda, MD, and an associate professor in the Department of Linguistics and the Institute for Advanced Computer Studies at the University of Maryland.

Michael Niv, PhD, is lead software architect at CodeRyte, Inc., in Bethesda, MD.

Michael Nossal, MA, is a senior NLP engineer at CodeRyte, Inc., in Bethesda, MD.

Andrew Kapit is CEO of CodeRyte, Inc., in Bethesda, MD.

Richard Toren is co-founder and president of CodeRyte, Inc., in Bethesda, MD.

Acknowledgments

The authors thank Lyle Schofield, the participating cardiologist, and an anonymous reviewer for their helpful comments.

Notes

1. Kaiser, Frieda, James Angus, and Helen Stevens (Editors). *e-MS Clinical Document Architecture Implementation Guide*. Government of British Columbia, 2006.
2. Cleveland, Harland. "Information as a Resource." *The Futurist*, December 1982, pp. 36–37.
3. Trachtenbarg, D. "EHRs Fix Everything—And Nine Other Myths." *Family Practice Management* 14, no. 3 (2007, March): 26–32.
4. It is not entirely clear whether this potential is realized using current EHRs. See Linder, J., J. Ma, D. Bates, B. Middleton, and R. Stafford. "Electronic Health Record Use and the Quality of Ambulatory Care in the United States." *Archives of Internal Medicine* 167 (2007): 1400–1405.
5. Trachtenbarg, D. "EHRs Fix Everything—And Nine Other Myths."
6. Waegemann, C. P., C. Tessier, A. Barbash, et al., for the Consensus Workgroup on Health Information Capture and Report Generation. *Healthcare Documentation: A Report on Information Capture and Report Generation*. Boston, MA: Medical Records Institute, June 2002.
7. Trachtenbarg, D. "EHRs Fix Everything—And Nine Other Myths."
8. For an informal user-satisfaction survey among 408 family physicians, see Adler, K.G., and R. L. Edsall. "An EHR User-Satisfaction Survey: Advice from 408 Family Physicians." *Family Practice Management* 12, no. 9 (2005, October): 29–35.
9. Shekelle, P. G., S. C. Morton, and E. B. Keeler. *Costs and Benefits of Health Information Technology. Evidence Report/Technology Assessment No. 132.* AHRQ Publication No. 06-E006. Rockville, MD: Agency for Healthcare Research and Quality, April 2006.
10. A sequence of 20 continuous notes was chosen starting at a random point in the collection. No two notes concerned the same patient, and no more than three notes were dictated by the same physician.
11. Kaiser, Frieda, James Angus, and Helen Stevens (Editors). *e-MS Clinical Document Architecture Implementation Guide*.
12. Guidelines for annotation were developed by independently marking up held-out development notes, comparing the results, and augmenting/revising the guidelines. There is currently vigorous effort among U.S. EHR standards bodies (the American Health Information Community and the Healthcare Information Technology Standards Panel) to define specific EHR standards. As soon as such a standard emerges, we plan to replicate this study using it, further extending our experimentation with a larger sample size and notes from other medical specialties. One annotator is a coauthor on this paper.
13. Words on a standard "stoplist" (*the*, *and*, etc.) were considered irrelevant, so highlighting *pain* and *neck* in the phrase *pain in the neck* would be considered equivalent to highlighting the full phrase.
14. The latter is a coauthor on this paper.
15. Kaiser, Frieda, James Angus, and Helen Stevens (Editors). *e-MS Clinical Document Architecture Implementation Guide*.
16. These three categories are neither mutually exclusive nor exhaustive. They illustrate patterns we found when we attempted to identify omissions that could be remediated by extending the EHR in obvious ways.

17. Grishman, Ralph, and Beth Sundheim. "Message Understanding Conference—6: A Brief History." *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, I, Copenhagen, 1996, 466–471.
18. Hripcsak, G., C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson, and P. D. Clayton. "Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing." *Annals of Internal Medicine* 122, no. 9 (1995, May 1): 681–688.
19. Resnik, P., M. Niv, M. Nossal, G. Schnitzer, J. Stoner, A. Kapit, and R. Toren. "Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation in Computer-assisted Coding." *Perspectives in Health Information Management*, Computer-assisted Coding Conference Proceedings, Fall 2006.
20. Engeler, C., J. Tashjian, S. Trenkner, and J. Walsh. "Ground Glass Opacity of the Lung Parenchyma: A Guide to Analysis with High-Resolution CT." *American Journal of Roentgenology* 160 (1993): 249–251.
21. Trachtenbarg, D. "EHRs Fix Everything—And Nine Other Myths."

Table 1

	Both experts			**MD cardiologist**		
Rating	□	□ □	□ □	□	□ □	□ □
Percent of notes	0%	50%	85%	55%	85%	100%
Number per note	0	0.6	1.85	1.45	3.45	5.25

Table 2

	Both experts			**MD cardiologist**		
Rating	**5**	□. □	□. □	**5**	□. □	□. □
Percent of notes	0%	25%	50%	45%	80%	95%
Number per note	0	0.3	0.95	0.75	2	2.65