AMA Journal of Ethics®

March 2017, Volume 19, Number 3: 281-288

STATE OF THE ART AND SCIENCE

Language, Structure, and Reuse in the Electronic Health Record

Angus Roberts, PhD

Abstract

Medical language is at the heart of the electronic health record (EHR), with up to 70 percent of the information in that record being recorded in the natural language, free-text portion. In moving from paper medical records to EHRs, we have opened up opportunities for the reuse of this clinical information through automated search and analysis. Natural language, however, is challenging for computational methods. This paper examines the tension between the nuanced, qualitative nature of medical language and the logical, structured nature of computation as well as the way in which these have interacted with each other through the medium of the EHR. The paper also examines the potential for the computational analysis of natural language to overcome this tension.

Introduction

The past few decades have seen a shift away from paper-based medical records towards computerized electronic health records (EHRs). Whereas paper-based records had their roots in a largely textual representation, the digital nature of computers lends itself more readily to the structuring and organization of data. The shift to the EHR has therefore been accompanied by a pressure on clinicians to record patient information in a structured way by choosing options such as diagnosis, medications, and symptoms from lists and completing onscreen forms. Structured information is computationally tractable, unlike the natural language of the textual portion of the record. Structured information, it is argued, can be reused to support research, audit, and the clinical process [1].

Very few would argue against the reuse of medical data. From the mid-sixteenth century, physicians increasingly recorded their cases, often indexed or ordered by disease or cure, in order to reuse them as a record of their practice and to extend medical knowledge. Thomas Willis, the seventeenth-century neuroanatomist, wrote that he would "weigh all the symptoms, and to put them, with exact Diaries of the Diseases, into writing; then diligently to meditate on these, and to compare some with others; and then [begin] to adopt general Notions from particular Events" [2].

Computer technology magnifies the efforts of Willis by many degrees, giving us the potential for reuse at scale. Structuring data allows the computer to aggregate, generalize, classify, sort, and search—powerful tools in building medical knowledge. We can imagine Willis leafing through his diaries to find a remembered patient, while the modern data analyst calls up 100 such cases. Whereas Willis could review his notes and compare one patient with another to find a pattern, a modern EHR-based study can crunch through tens of thousands of records to find small but statistically significant relationships [3].

There is, however, a problem. The EHR exposes a fundamental conflict between the needs of software and the needs of human users. The EHR tries to bridge two worlds: the human, "analogue," cognitive world and the formal, logical, "digital" world of the machine [4]. There are many ways in which EHR design tries to overcome this conflict and bridge the analogue and digital worlds. I examine some of these below and argue that such designs fail to capture a full record of the patient. This leaves clinicians falling back on recording clinical encounters in analogue, through the use of natural language text. If we are to reuse the data of the EHR, then we must find ways to analyze this text. I look at how natural language processing—the computational analysis of natural language text—offers a way to do this.

Bridging the Digital World of the EHR and the Analogue World of the Clinician

One attempt to bridge the analogue and digital worlds can be seen in the use of medical terminologies in the EHR. Such terminologies are not intended to replace clinical narratives but rather to allow the coding of events alongside the narrative text of the record. In their simplest form, these are lists of codes, each associated with a human language term for some disorder or class of disorder, often arranged in taxonomies. A simple terminology, however, no longer satisfies the needs of administrative coding, leading to the introduction of ever more complex terminologies. This problem is illustrated in the following description of the ICD-10 terminology, recently introduced in the United States. "Coding for medical encounters used to be haphazard" says the author, but this will change as "ICD-10 has a new structure and more room (up to seven characters, from five)" and includes "details, such as laterality and etiology" [5]. Such a coding scheme allows for grouping and analysis of clinical encounters, but we would clearly need other techniques if we wanted to capture all the detail of those encounters in structured form. While ICD-10 may be more powerful at coding than its predecessor, seven characters and a few added details are never going to have the expressivity of even a single sentence of natural language. Coding, both intentionally and as a result of the limits of what can be practically described, is about generalizing. While coding schemes might accurately describe a patient as a member of some group, they were never designed to describe the individual patient in detail.

Accordingly, there have been several efforts to provide ways of structuring the record of the clinical encounter. In computer-based documentation (CBD) systems, documentation is driven by the completion of onscreen templates: picking items for diagnosis, symptoms, interventions, and medications from drop-down boxes; check lists; and other computer interface components. The selection of the appropriate templates for completion might be driven by computerized workflows and care pathways [6, 7]. For example, entering that a patient smokes may lead to questions about how many cigarettes per day are smoked and for how long the patient has been smoking. Or recording a specific test result may prompt the user to consider other investigations. In the closely related structured data entry (SDE) approach, the user creates documentation by selecting clinical concepts from interfaces constructed from some underlying knowledge model, usually based on standard medical terminologies. Concepts may be further qualified and adapted by selecting modifiers, anatomical location, and so on [8]. Selecting "abdominal pain," for example, may lead to the user being given a choice of more precise localization and a choice of onset.

SDE and CBD may well provide a rich and convenient way of describing patients. Yet, however data are structured, the clinician can only consider the fixed set of patient characteristics and features allowed by the structured representation and has no way to stray beyond those parameters. Clinicians cannot easily describe the personal, social, and cultural circumstances of patients; the interplay between their disease, life, and treatment; or the particular way in which they experience their disease. Nor can clinicians give a detailed description of the clinical encounter and of their personal reaction to it. Instead, patients are described as members of a population that share the same limited structured representation. Swinglehurst, talking about the UK primary physician record, which has been highly structured since the 1990s, describes a "dilemma of attention" [9]. On the one hand, medicine frames the patient as an individual and, on the other, as part of a population. The EHR brings this dilemma into sharp focus with easily structured and easily coded "hard" data pushing the dilemma's resolution towards framing the patient as part of a population and representing an increase in the bureaucratization of health care. Should patients be treated only as members of populations, or is there some value in considering them as individuals? And, if we need to consider them as individuals to give the best care for their circumstances, is structured data able to convey all of the information necessary to support this care?

Analyzing the Text of the EHR

Despite the efforts put into structuring the clinical narrative, the fact that structured representations are not able to give the level of description and convenience required by the clinician means that the medical record is still dominated by unstructured natural language. While CBD and similar ideas have a place in many EHR systems, the addition of free-text notes and the uploading of documents remain common EHR functionalities. Indeed, many important observations go unrecorded in the structured record, only

appearing in the free text stored alongside the empty fields and forms. In one UK case register derived from a forms-based EHR, for example, dealing with free text has been a major concern of reuse [10].

Why do clinicians prefer text and insist on using it? Meystre et al. note that free text is convenient to express clinical concepts and events, such as diagnosis, symptoms, and interventions [11]. Reviewing the few studies that look at the expressivity of CBD systems compared to natural language notes, Rosenbloom et al. report that prose can be more accurate, reliable, and understandable [12]. Powsner, Wyatt, and Wright refer to structured data as freezing clinical language and restricting what may be said [13]. Much of medical language is nuanced and makes heavy use of negation, temporal expressions, and hedging phrases. These are all difficult to represent as structured data. For example, when saying that something happened "a few months ago," or that it is "more or less resolved," the time and resolution cannot easily be accommodated by structured elements. Greenhalgh et al. say that free text is tolerant of ambiguity, which supports the complexity of clinical practice [14].

One way in which this tension may be resolved is through a linguistic analysis of the free text: an area of computer science known as natural language processing (NLP). NLP of medical records is nearly as old as the computerization of those records. Sager's Linguistic String Project, for example, implemented NLP in radiology reports in 1976 [15]. The last few years have seen a big growth in medical NLP—paralleling the growth in the EHR—stimulated by government investment in health information technology (IT) internationally. Uses include automated coding of episodes, extraction of facts such as symptoms and confounding factors to support epidemiology, and extraction of clinical events to drive decision support [11, 16].

NLP does not provide a complete answer to the problem of extracting information from natural language, though, as the very reason clinicians value language—its expressivity—makes it difficult to analyze. The challenges that NLP faces are technical, organizational, and social. Specific technical challenges for NLP include ambiguity, uncertainty, complex temporal reasoning, complex terminology, heavy use of abbreviations, and a wide range of texts from prose-like letters to terse reports. All of these are active areas of NLP research [11]. There are also social and organizational challenges to its adoption. For example, the development of an NLP system usually requires example texts with the phenomenon of interest already identified, for the purpose of both training by example and evaluation. Marking the phenomenon of interest in the data requires expert human resources, often scarce in a health setting. Moreover, exporting the data from its source EHR and sharing it with software developers and the NLP research community raises privacy issues. And, finally, after all the effort, the NLP system will still make mistakes. The end user—perhaps a data

analyst, perhaps a clinician using a decision support tool—has to deal with the system's inevitable errors.

Much of medical NLP is targeted at extracting quantifiable facts expressed directly in the text, such as finding test results and medications discussed in an encounter note or finding a patient's symptoms and smoking status from a clinic letter. We may, however, go beyond extraction of bare facts from individual records and study variation in the corpus as a whole, finding information that the writer may not have consciously intended to reveal. For example, McCoy et al. [17] studied sentiment expressed in discharge notes by looking at occurrences of words related to polarity (positive or negative), subjectivity, intensity, and negation. They found that, for psychiatric patients, public insurance was associated with significantly lower levels of positive sentiment while greater comorbidity was associated with significantly lower levels of both positive and negative sentiment. Additionally, self-identification as Hispanic was associated with significantly higher levels of both positive and negative sentiment. A similar approach has been used to study suicide risk, with one study finding that the clinic notes of outpatients who later died from suicide showed an increase in distancing language—for example, an increase in the use of third-person pronouns by the clinician [18].

These examples expose the power of natural language communication and give an insight into why clinicians value it. There is a sense in which the language of the record—particularly the narrative parts such as letters between clinicians—carry more information than could ever be conveyed by structure alone. When physician and anthropologist Cecil Helman describes reading a "fat file … filled with the frustrated letters of a dozen doctors" and goes on to talk about the "tone" of those letters and the "hints" they contain [19], he is describing how communication through narrative text goes beyond a stream of facts. Through natural language, we communicate thoughts and feelings that we may only be dimly aware of ourselves. The implication is that by getting rid of the natural language text of the EHR, we will remove that communication and all of its benefits.

Conclusion

While structuring EHRs is a valuable way to bring benefit by allowing their reuse, we also need to recognize the importance of natural language in human communication and allow for it when building EHRs and when deploying technologies, such as NLP, to analyze those EHRs [12].

What will the future EHR bring to the language of the medical record? One possibility is that records will no longer be confined to communication between clinicians and that patients will join in the conversation. Legislation now allows patients to see their records in many countries. For example, in the UK, this is enshrined in the Data Protection Act [20]. With the buff folder hidden away in a basement medical records library, it was

impractical for patients to regularly review their own record. Unlike paper records, however, the EHR is instantly portable and can be viewed at any location and at any time. We might expect that patients will increasingly access their own notes by browsing them on the web or swiping through them on their phone. This accessibility has implications for how records should be best presented to patients in order to aid their understanding and to avoid unnecessary alarm. It is as yet unclear how patient access will change the way in which physicians interact with the record, although there is evidence that, once again, physicians will not feel the need to change their language [21].

References

1. van der Lei J. Closing the loop between clinical practice, research, and education: the potential of electronic patient records. *Methods Inf Med*. 2002;41(1):51-54.
2. Willis T. *Dr. Willis's Practice of Physick Being All the Medical Works of That Renowned and Famous Physician*. Pordage S, trans. London, England: T Dring, C. Harper, and J. Leigh; 1681:55. Quoted by: Kassell L. Casebooks in early modern England: medicine, astrology, and written records. *Bull Hist Med*. 2014;88(4):614.
3. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. *Yearb Med Inform*. 2014;9(1):97-104.
4. Rector A. Clinical terminology: why is it so hard? *Methods Inf Med*. 1999;38(4-5):239-252.
5. Rubenstein J. ICD-10: are you ready? *Curr Urol Rep*. 2014;15(11):449.
6. Johnson KB, Ravich WJ, Cowan JA Jr. Brainstorming about next-generation computer-based documentation: an AMIA clinical working group survey. *Int J Med Inform*. 2004;73(9):665-674.
7. Gooch P, Roudsari A. Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *J Am Med Inform Assoc*. 2011;18(6):738-748.
8. Bleeker S, Derksen-Lubsen G, van Ginneken AM, van der Lei J, Moll HA. Structured data entry for narrative data in a broad specialty: patient history and physical examination in pediatrics. *BMC Med Inform Decis Mak*. 2006;6:29. http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-6-29. Accessed January 27, 2017.
9. Swinglehurst D. *The Electronic Patient Record: A Linguistic Ethnographic Study in General Practice* [dissertation]. London, England: Queen Mary University of London; 2012.
10. Perera G, Broadbent M, Callard F, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an electronic mental health record-derived data

resource. *BMJ Open*. 2016;6(3):e008721. http://bmjopen.bmj.com/content/6/3/e008721.long. Accessed October 28, 2016.
11. Meystre S, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128-144.
12. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc*. 2011;18(2):181-186.
13. Powsner SM, Wyatt JC, Wright P. Opportunities for and challenges of computerisation. *Lancet*. 1998;352(9140):1617-1622.
14. Greenhalgh T, Potts HW, Wong G, Bark P, Swinglehurst D. Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. *Milbank Q*. 2009;87(4):729-788.
15. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc*. 1994;1(2):142-160.
16. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform*. 2015;10(1):183-193.
17. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. *PLoS One*. 2015;10(8):e0136341. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136341. Accessed October 28, 2016.
18. Leonard Westgate C, Shiner B, Thompson P, Watts BV. Evaluation of veterans' suicide risk with the use of linguistic detection methods. *Psychiatr Serv*. 2015;66(10):1051-1056.
19. Helman C. The other half of Eddie Barnett. In: Helman C, ed. *Doctors and Patients: An Anthology*. Oxford, UK: Radcliffe Medical Press; 2003:124.
20. Data Protection Act 1998. http://www.legislation.gov.uk/ukpga/1998/29/contents. Accessed December 24, 2016.
21. Kind EA, Fowles JB, Craft CE, Kind AC, Richter SA. No change in physician dictation patterns when visit notes are made available online for patients. *Mayo Clin Proc*. 2011;86(5):397-405.

Angus Roberts, PhD, is a senior research fellow at the University of Sheffield, Sheffield, United Kingdom. He also leads life science-related work for GATE, a widely used open-

source platform for large-scale text mining and language engineering. His research is in the area of medical informatics, with an emphasis on deployment in real-world settings.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect).

Related in the *AMA Journal of Ethics*

Development of the Electronic Health Record, March 2011
Does Health Information Technology Dehumanize Health Care?, March 2011
Electronic Health Records: Privacy, Confidentiality, and Security, September 2012
Mindful Use of Health Information Technology, March 2011
When and How Should Clinicians Share Details from a Health Record with Patients with Mental Illness?, March 2017

ISSN 2376-6980