Medical Hypotheses 92 (2016) 75–79

Contents lists available at ScienceDirect

Medical Hypotheses

journal homepage: www.elsevier.com/locate/mehy

The lexeme hypotheses: Their use to generate highly grammatical and completely computerized medical records

Donald Macfarlane *

Department of Internal Medicine, The University of Iowa, United States

ARTICLE INFO

Article history:
Received 15 March 2016
Revised 12 April 2016
Accepted 16 April 2016

ABSTRACT

Introduction: Medical records often contain free text created by harried clinicians. Free text often contains errors which make it an unsuitable target for computerized data extraction. The cost of healthcare can be reduced by creating medical records that are fully computerized at their inception. We examine hypotheses that enable us to construct such records.

Methods: We regard the text of the medical record as being an ordered collection of meaningful fragments. The intellectual content (or "lexeme") of each text fragment in the record is considered separately from the language that used to express it. We further consider that each lexeme exists as a combination of a lexeme query (defining the issue being addressed) and a lexeme response to that query. The medical record can then be perceived as a stream of these responses. The responses can be expressed in any style or language, including computer code. Examining medical records in this light gives rise to a number of observations and hypotheses.

Observations and Hypotheses: The physical location and nature of the medical episode (which we term "context") determines the general layout of the record. The order that lexeme-queries are addressed in within the record is highly consistent ("coherence"). Issues are only addressed if they are logically called-for by the context or by a previously-selected lexeme response ("predicance"), and only to a needed depth of detail ("level").

We hypothesize that all of the lexeme queries required to write any clinical notes can be stored in a large database ("lexicon") in coherence order, wherein each lexeme query is associated with its own collection of lexeme responses. We hypothesize that the issue a note-writer will need to address next is identifiable purely by using the rules of coherence, level and predicance.

Testing the hypotheses and their utility: We have tested these hypotheses with a computer program which repeatedly offers the user a menu of lexeme responses with associated text. On selection, the program issues the text fragment, and its corresponding computer code, to output files. The program then uses coherence, predicance and level to navigate to the next appropriate lexeme query for presentation to the user. The net result is that the user creates a grammatically correct and completely computerized note at the time of its inception. The value of this approach and its practical implementation to create medical records are discussed.

In our work so far, the hypotheses appear not to be false, but further testing is needed using a larger lexicon to establish their robustness in actual clinical practice.

Statement of the problem

A century and a half ago, Florence Nightingale urged hospitals to issue a report as to how well they cared for patients [1]. Her plea was echoed by the Boston surgeon, E A Codman, whose "End Result Idea" urged hospitals to take the (then-radical) step of taking ownership of patients' case notes and analyze them, to enable the public to assess the hospital staff's performance [2].

Codman and Nightingale hoped that the practice of medicine could be improved by comparing the outcome of aggregated case notes. Their dream is yet to be fulfilled. Over the intervening century, the creation and analysis of patient's medical notes have become vastly more inefficient. And vastly more expensive: In the US, creating notes occupies about 15% of physician time, and employs 184,000 dedicated personnel [3]. The storage and distribution of medical records drive the need for hospital information systems costing $33 billion annually [4], and electronic medical record systems costing $18 billion annually [5]. Health insurers use these notes to determine whether or not they will pay

* Address: Internal Medicine, University of Iowa, Iowa City 52242, United States.
E-mail address: donald-macfarlane@uiowa.edu

http://dx.doi.org/10.1016/j.mehy.2016.04.031
0306-9877/

submitted bills, transactions that add about $210 billion annually to the cost of health care delivery [6]. The creation and manipulation of medical records therefore account for a staggering 2.5% of the nation's gross domestic product.

Despite this vast economic activity, there is a paucity of academic research concerning the medical record. Instruments developed and validated to judge how well a note is written reveal a wide variation in perceived quality [7]. 60% of inpatient notes contain errors, and those that do average 7.8 errors per patient. Physicians express widespread dissatisfaction with electronic medical record systems [8]. The most effective way to write a note is not established by clinical trial.

There is no agreement within the profession as to how a note should be ordered, such as where the psychiatric or lymph node exam should be placed, whether the respiratory exam should precede or follow the cardiac exam [9,10], and whether notes should be linearly arranged coincident with the typical history and physical exam, or organized in a problem-oriented format [11,12]. Casual observation supports the view that individual clinicians develop their own unique means of recording their interactions with patients.

The modern electronic medical record has enabled bloated notes to be created, stored and disseminated – complete with factual, typographical and cut-and-pasted errors. It has, at its heart, narrative text generated by idiosyncratic and harried practitioners, who feel free to express their findings in a wide variety of choices of language. For instance, in a study of how to address possible drug interactions, clinicians at one institution were required to respond by selecting a response from a menu. At a similar institution, clinicians were required to respond by entering free text. Clinicians found 209 different free text ways to express the thought corresponding to the menu item "will monitor as recommended" [13]. We need to understand how this diversity of text to address a simple issue arises.

Homo sapiens is endowed with what Chomsky termed a "language acquisition device", enabling each of us to generate and express language [14]. It also enables us to extract meaning from other people's language despite idiosyncracies and errors. The clinician writes notes with the expectation that the reader has well-developed acquisition device. As a result, the clinician feels little restraint on the language used to create the text, and feels no great pressure from colleagues to correct even the most obvious errors.

Our language skill enables the flow of information illustrated in Fig. 1. The clinician formulates a thought and expresses it, using a wide diversity of error-prone language. The reader uses language skill to abstract the clinician's original thought, intuitively correcting some of the errors.

A substantial investment of effort and resources has been made to automate the extraction of the writers' original thoughts from the free texts they created. This effort has led to only modest success (reviewed here [15]). Even trying to extract a simple time-line of a sequence of events described in typical free text is difficult and error prone [16,17].

To limit the use of free text, a variety of attempts have been made to enable clinicians to enter some computer-readable clinical information, offering a hybrid between structured data interspersed with narrative text, eg [18]. These templated, or menu-driven systems, may work well in simple and repetitive areas of practice, but most clinicians feel the need to use fully nuanced text to express their findings adequately.

The fact remains that accurate analysis of the content of any medical record today can be only achieved if it is read carefully by a knowledgeable individual. This adds vastly to the difficulty and expense of all aspects of healthcare delivery, research and administration.

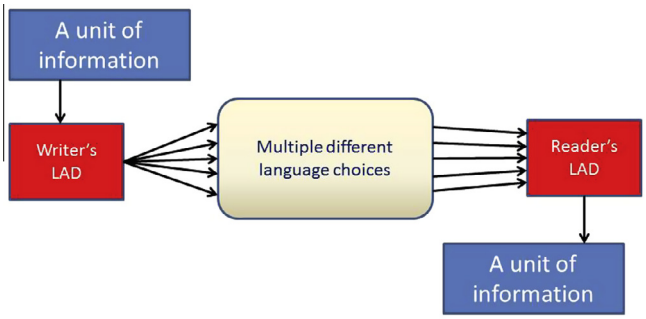

Fig. 1. A diagram of the logic flow of current practice. The creator of the note formulates a unit of information, and then uses the language acquisition device to create any one of a large number of language choices to express it. The reader uses the language acquisition device to convert the language back to the unit of information.

Generating a theoretical basis

In this article, I analyze the process clinicians use to construct medical records. This analysis is undertaken to provide a theoretical basis for automatically generating medical notes that are fully nuanced, grammatically correct, and completely computer readable at the time of their creation. The approach we use gives rise to the flow of information illustrated in Fig. 2, in which the note writer selects one appropriate language choice to express a thought, instead of creating new text.

Lexeme

The first step in the analysis is to use the device of separating the *meaning* of a fragment of text from the *language* used to expresses it. We will refer to this meaning as a lexeme[1]. Using this analytical device, we can regard the medical record as consisting of an ordered collection of lexemes (facts or opinions), each expressed as a fragment of text. The typical medical record might contain a few dozen lexemes. A complete collection of all the lexemes that might be required to provide all the language needed by all the branches of medicine with an adequate degree of subtly would likely run in the millions of items.

Lexemes correspond with the formulated thoughts of Fig. 1. Each lexeme can be expressed in any *style* (eg courteous, abbreviated) or language – including (to great advantage) in computer code[2].

Lexeme query

The second step in the analysis is artificially to regard each lexeme as being a composite of a question or *lexeme query*, which effectively defines the issue being addressed, and an answer, or *lexeme response*, which addresses that specific question[3]. Within this analysis, each lexeme response must be associated with, and therefore be additionally defined by, a lexeme query. To illustrate the significance of this, consider the response expressed as "the pain is getting worse". This response has a totally different clinical implication when the lexeme query associated with it addresses cardiogenic chest pain than it does when addressing a gouty big toe.

[1] *Lexeme* is defined as a the least unit of information, divorced from the text used to express it.

[2] *Style* refers to the language choices that are made to express the information in a lexeme.

[3] *Lexeme query* refers to the issue that addressed by a lexeme, and lexeme response refers to the (limited number) of answers to query.