Chapter 7.17
Preparing Clinical Text for Use in Biomedical Research

John P. Pestian
Cincinnati Children's Hospital Medical Center, University of Cincinnati, USA

Lukasz Itert
Nicolaus Copernicus University, Torun, Poland

Charlotte Andersen
Cincinnati Children's Hospital Medical Center, University of Cincinnati, USA

Wlodzislaw Duch
Nicolaus Copernicus University, Torun, Poland

ABSTRACT

Approximately 57 different types of clinical annotations construct a patient's medical record. These annotations include radiology reports, discharge summaries, and surgical and nursing notes. Hospitals typically produce millions of text-based medical records over the course of a year. These records are essential for the delivery of care, but many are underutilized or not utilized at all for clinical research. The textual data found in these annotations is a rich source of insights into aspects of clinical care and the clinical delivery system. Recent regulatory actions, however, require that, in many cases, data not obtained through informed consent or data not related to the delivery of care must be made anonymous (as referred to by regulators as harmless), before they can be used. This article describes a practical approach with which Cincinnati Children's Hospital Medical Center (CCHMC), a large pediatric academic medical center with more than 761,000 annual patient encounters, developed open source software for making pediatric clinical text harmless without losing its rich meaning. Development of the software dealt with many of the issues that often arise in natural language processing, such as data collection, disambiguation, and data scrubbing.

INTRODUCTION

Hospitals typically produce millions of text-based medical records over the course of a year. These

records are essential for the delivery of care but underutilized or not utilized at all for clinical research. Digitized clinical data are a rich lode of possibilities for advances in biomedical research, because, in aggregate, they contain information about the variation in the delivery and quality of care.

Inherent in such research, however, is the use of data without the patient's consent. Recognizing this problem, the United States Department of Health and Human Services (HHS) has issued rules defining Protected Health Information (PHI) as part of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (Annas, 2002). In order for researchers to access such data, either they must have the patient's consent, or, as in most retrospective cases, the data must be made harmless, and the governing board must provide a waiver.

The HHS provides guidance for making healthcare data harmless (HIPAA Standards for Privacy of Individually Identifiable Health Information: An Introduction to the Consent Debate, 2002). Data can be made harmless through three steps: (1) deidentification (i.e., the removal or modification of data fields that could identify a patient, such as name and social security number); (2) rendering the data ambiguous by ensuring that every data record in a public data set has a non-unique set of characterizing data (Berman, 2002a; Bouzelat, Quantin, & Dusserre, 1996; Quantin et al., 1998); and (3) data scrubbing (i.e., the removal or transformation of those tokens in text that can be used to identify persons or that contain information that is incriminating or otherwise private) (Berman, 2003; Sweeney, 1996). Although each of these methods has the potential to render the medical record harmless for its use by natural language processing investigators, attempts to design a fully anonymous system continue.

This article describes how Cincinnati Children's Hospital Medical Center (CCHMC), a large pediatric academic medical center with more than 761,000 pediatric patient encounters per year, has taken a practical approach to this challenge by developing, evaluating, and implementing the Encryption Broker (EB) software. The EB has a number of uses. First, it is essential for the ongoing development of a large pediatric corpus for pediatric natural language processing research and decision support (Pestian, Itert, & Duch, 2004). This corpus serves as an artificial intelligence training set for classifying text into the appropriate clinical domain, such as rheumatology or neonatology. Without the EB, these data could not be retrieved from the electronic portion of the medical records. Second, the EB ensures that research-needing text conforms to federal regulations. It does so through data disambiguation algorithms, deidentification, and data scrubbing.

The EB has another role. A key strategy of the organization is personalized medicine research that requires genomic and clinical delivery data to predict or prevent disease or to personalize treatment. This research requires substantial amounts of knowledge to be gleaned automatically from these data in real time. To do so, machine-learning systems that conceptually map the data into some ontology are required. The EB provides natural language scientists with large repositories of harmless clinical text for developing these systems.

The EB is recognized by CCHMC's Risk Management group as a tool to gather clinical text without violating HIPAA regulations. This approval is institution-specific; each institution using the EB is responsible for seeking its own internal certification. The EB essentially acts as a broker for investigators who wish to do retrospective analysis of clinical text and potentially makes it easier to receive approval for these purposes. CCHMC makes the EB software, the associated decision rules, and the related data files fully available for academic purposes by e-mailing joan.taylor@cchmc.org. The remaining sections of this article discuss methods and challenges for