# Electronic Health Records:

## Sharing knowledge while preserving privacy

Catherine Arnott Smith, Ph.D.
Assistant Professor
School of Information Studies
Syracuse University
Syracuse, NY

## Table of Contents

# Introduction

## Overview

After decades of discussion, electronic health records finally are beginning to come into their own. The National Health Information Infrastructure (NHII), a federal effort launched in mid-2004, is leading the effort to standardize information for exchange industry-wide. Progressive, information-savvy healthcare providers such as Mayo Clinic and Kaiser Permanente are using and refining their own versions of electronic health records (EHRs). One healthcare industry study even rated EHRs second on its list of the "Top Ten Trends for the Future," placing them second only to patient safety as an executive focus in 2010.

Only a small minority of institutions currently use EHRs, but pressure for their implementation is building rapidly. Simply put, EHRs make sense for diagnostic, treatment, research and business reasons. By converting disorganized paper files to electronic versions, physicians and hospitals are able to access information far more quickly, and care providers are less likely to overlook information crucial to an individual's health and treatment. Making records available via intranet to multiple, authorized users simultaneously eliminates the risk that errors will result from outdated information, even if that information is only a few minutes old. All these factors meaningfully can improve patient safety and quality of care.

By enabling healthcare workers to do their jobs rather than waste time simply trying to locate patients' charts or diagnostic images, EHRs enable hospitals and other institutions to operate far more cost-effectively. EHR availability also suits an increasingly mobile population, which may seek out care at hospitals, outpatient centers, mobile facilities, or by phone. Savvier consumers also want greater access, including online access, to their own health- and cost-related medical records. And standardized formats and language could make EHRs far more useful for medical research than current paper systems, whose sheer disorganization can be a substantial disincentive for analysis.

Yet EHRs aren't yet a *fait accompli*. While a great deal of attention has been paid to what types of information such records should contain, relatively little has been written about how that data should be structured. Organization is critical for several reasons. In clinical care, standard formats and data types clearly would enable physicians and other experienced users to find the information they seek most quickly. In research, standardization of categories and information within those categories renders EHRs far more searchable for epidemiology, genetics or numerous other public health purposes, including data mining for valuable patterns that otherwise would remain undiscovered. And in administration, structure of information could aid greatly with discerning expenditures of time, supplies and funds that may be managed more effectively.

Most pressingly, however, is the need to comply with privacy regulations in the Health Information Portability and Accountability Act of 1996, commonly referred to as HIPAA. Under HIPAA, healthcare institutions must meet strict confidentiality requirements for maintaining medical records, whether electronic or paper-based, and for exchanging the clinical data they contain between healthcare providers. (The exception is when records are required for treatment purposes.) These standards require stripping numerous types of specific information about a patient, as well as any relatives or other individuals whose medical data may be included, from records before access is permitted for research or other non-treatment-related purposes. This process is known as "de-identifying" records.

In short, EHRs create a research paradox. Health information is most valuable when it's shared. Yet sharing creates the risk of exposing highly sensitive information about specific individuals, even family members or others who are not patients themselves. Only by understanding what information is contained in medical records can we decide what the best structure for EHRs will be – in terms of both accessibility and security.

This paper explores the evolution of EHRs, their ideal form and functions, and major constraints inhibiting that development to date. Using records of patients with highly complex, chronic medical conditions, which accordingly require extensive documentation, the paper investigates the substance and structure of information as collected by a representative small urban hospital. It then examines that information from the perspective of the Clinical Document Architecture (CDA), an emerging standard being developed to accommodate diverse medical records and requirements. Conclusions include proposals to maximize application of the CDA when converting such disparate, yet crucial, collections of paper-based and existing electronic records to the next generation of medicine: the EHR.

### Study Recommendations

The Clinical Document Architecture (CDA) is an emerging U.S. standard receiving international recognition. It has the potential to address the conflicting aims of ease of information access and protection of patient privacy through effective document modeling for clinical information exchange. To fully exploit the CDA, however, requires an understanding of the clinical documents that make up the medical record today.

Review of typical paper records in a small urban hospital shows large numbers of unique documents with relatively standardized and recurring data elements, which are restricted from release under HIPAA without expensive and intensive de-identification. In addition, more than half of the unique document types appeared in extremely low frequencies. Analysis of data types routinely collected — for example, upon patient registration and intake into the health care system — should assist in developing a data model for electronic data exchange under national healthcare information standards. Considerable efficiencies simultaneously could be realized through internal analysis of the degree of repetition and overlap between these unique documents. The result would be enhanced value of existing data, for research and treatment purposes, as well as decreased vulnerability to HIPAA violations.

## Background

### History of the Medical Record

The medical record has existed as a feature of patient care from the time of the Codex Hammurabi, the legal system named for a king of Babylon who reigned somewhere between 2067 and 1169 BC. According to Spiegel and Springer (1997), "Thousands of clay tablets recovered by archaeologists document that medical care data were collected and recorded about ailments, causes, treatments, and therapy outcomes[1]."

Four principal ancestors of the modern medical record have been identified, and the modern record resembles all of these in substance, as well in function. They are: (1) the case record collection of the 19th century, resembling "diaries or research notebooks"; (2) the bedside chart, on which an individual patient's vital signs and observations were recorded; (3) the physician order used for communication of directives about patient care to staff members; and (4) the financial ledger in which physician charges and transactions were recorded[2]. At present, the vast majority of medical records are kept either entirely on paper (and, in the case of certain images, on film), or in a combination of paper and electronic forms. The typical modern hospital's healthcare information system contains islands of proprietary systems, each dedicated to the needs of one typical department. The "best of breed" mentality prevalent in healthcare IT ensures that department-specific applications take priority over interoperability, thus reinforcing the existing disdain for standards.

The result has been a cafeteria of incompatible EHR software solutions. A 2003 survey of family physicians, for example, revealed 274 unique EHRs in use – only a few of which had more than one user. In contrast, healthcare experts at the Gartner Group have recommended that "the best approaches to clinical information technology allow data from perhaps dozens of different computer applications to be swapped and sorted as if they came from the same system".[3]

Hospitals and other healthcare providers are well aware that the medical records status quo presents a significant business and information management problem. The Medical Records area was identified as the most problematic department in a 2003 survey of Health Information Management Systems Society members, of whom 45% were chief information officers and 38% were directors of information technology. A full 75% spent less than $500,000 annually on document management, while 54% of all respondents acknowledged they didn't track such costs in detail.

Small wonder that the desire to automate medical records can be traced back several decades. Electronic health record construction was the driving force behind the development of hospital information systems. In 1991, the Institute of Medicine first called for a paperless system to be in place in the next 10 years. The task, however, proved highly resource-intensive. By 2001, those calling for a national health information infrastructure found that only 13% of healthcare providers had a fully operational system in place[4]. EHR penetration of 5% has been reported among primary care providers[5], up to 10% in the U.S. market nationally[6], and 23.5% among family physicians[7]. In short, despite the longstanding need for EHRs, the task clearly is easier visualized than done.

### About EHRs

The EHR has been defined both as a "secure, point of care, patient-centric information resource for clinicians[8]," and as a "complete online record that is accessible to all that need it when it is needed[9]". It has also been called a "container for a set of transactions." These include "persistent" transactions with long-term value, such as historical data pertaining to one patient, and "event" transactions with short-term value, such as EKG tracings of that one patient on one morning in the clinic.[10]

Along with EHRs, researchers and clinicians use an alphabet soup of other acronyms for the same suite of patient information records. These include clinical data repository (CDR), computerized patient record (CPR), electronic patient record (EPR), electronic health record (EHR), and electronic medical record (EMR.) As the term used by David Brailer, National Coordinator for Health Information Technology — in first articulating the NHII initiative in July of 2004 — EHR has been chosen for this paper.

One characteristic that distinguishes healthcare IT from that of other industries is the need for longitudinal

information access. Susan Cisco, currently project manager for consulting services at records storage and management specialist Iron Mountain Inc., has said that healthcare entities "have heavy retrieval requirements initially and then a drop-off in the need to access records … it is not unusual for a caregiver to need access to 20 years' worth of a patient's medical history ... there is no predictable retrieval pattern for medical records."[11] For the same reason, the Institute of Medicine, in its report Key Capabilities of an EHR System, stresses that the modern motivation for an EHR is not to have "a paperless record per se, but to make important patient information and data readily available and useable."[12]

Paper medical records historically have supported numerous work processes and sub-processes, often with multiple authors and custodians. Records also may have multiple audiences, intended data lifespans, and trajectories documenting care in different locations and/or for different purposes. Clearly, EHRs must continue to serve the same functions. Yet they also must support growing quantities of contemporary business-related information: records management, process management, outcome management, demand management, and health management.

The ideal functions supported by the EHR include these, outlined by HIMSS in its Definitional Model (2003):[13]

- Captures and manages episodic and longitudinal electronic health record information
- Functions as the clinician's primary information resource during the provision of patient care

Specifically, the EHR provides health information and data; results management; order entry management; decision support; electronic communication and connectivity; patient support; administrative processes; and reporting and population health management.

To support these important functions, EHRs typically contain a mix of highly structured numeric data and excessively unstructured and idiosyncratic narrative. In fact, any information relevant to clinical decision-making can be part of the medical record. This data makes its way into the record via voice transcription, data feed from machines, or conversion from paper. Although there is considerable variation in the content and structure of medical records, the current paper-based record contains these typical contents that must be simulated by the EHR:

- patient problem list
- patient history
- operating room notes
- physical exams
- discharge summaries
- allergies
- health maintenance information
- immunizations
- medications dispensed
- orders
- diagnostic results
- images
- most recent vital signs
- progress notes
- nursing visits
- consult documentation
- genetic information
- results of previous retrieval runs of any or all of the above,

and information that has been generated outside the healthcare organization. The latter may include content in all of the above categories. Other externally generated materials could take the form of letters from referring physicians that reference radiology report results, or from a physician to the patient or his or her family members discussing the impact of genetic testing.

### The Drive Toward EHRs

EHRs clearly are vital to the emerging electronic healthcare environment. Development of EHRs is recognized explicitly as a major goal of the proposed National Health Information Infrastructure (http://aspe.hhs.gov/sp/nhii/). The NHII is simultaneously a federally driven initiative, a network, and a set of technologies designed to enable healthcare information management in the United States. Financial incentives for EHRs also have been built into the Medicare Modernization Act of 2003.

Other pressures moving the healthcare industry quickly toward EHRs include:

- the prevalence of chronic disease, and thus of greater need for information management related to chronic disease;
- the need for timely access to information;
- the need for simultaneous access by multiple caregivers;
- multiple settings of healthcare information need, including mobile health, telemedicine, and telecare;
- increasingly mobile patient populations;
- better cost-effectiveness in an environment of rapidly escalating costs; and
- better support for clinical research.

Patients' and consumers' rising demands for access and participatory decision-making constitute another source of pressure. Consumers appear to support electronic information management of healthcare records, although concerns over security of healthcare data are likely to remain an issue. In one study, 34% of consumers polled reported they would pay extra to manage their benefits online, 25% would pay more for online interaction with physicians, and 25% would actually switch either insurers or physicians to be able to do so.[14] In another large study, 44% of policyholders expressed a preference for electronic access to their medical records.[15]

### The Business Case for Standards

One formidable barrier to implementation of fully operational EHRs has been quantifiable return on investment (ROI). Respondents to a recent HIMSS survey identify the business case for EHR adoption by physicians and hospitals as the issue having the most potential impact on implementation of these systems. Second to EHR adoption was patient safety, meaning the reduction of medication errors and other adverse events.[16] The business case for EHRs has proven elusive given that the healthcare industry already regards information systems simply as "additional cost".[17] Overall, hospitals and other providers invest only 2% of gross revenues in IT—a fifth of the amount dedicated by similarly information-intensive industries[18].

Lang, writing for the Journal of Healthcare Information Management's special ROI issue, notes that establishing ROI is a difficulty of IT projects in general. As Vogel states, "investments in information technology create an asset that is truly different from other assets that organizations have traditionally created and understood"[19]. It may be that traditional ROI calculations cannot account for qualitative aspects of healthcare work in particular. Such "soft returns" as improvements in customer relations, innovation, patient safety and internal business processes can be extremely hard to quantify. Thus, despite the considerable potential of EHR projects to financially benefit an organization, initiatives to digitize its patient records may stall or be abandoned[20].

Information has characteristics that distinguish it from other, more tangible business assets. Information is shareable, usable and reusable without decreasing in amount, characteristics that have implications and consequences for social as well as business processes. EHRs that constructively exploit these information properties can generate benefits in multiple realms, including:

- **Clinical results**. Improved adherence to treatment following patient education has been demonstrated repeatedly. Such education requires access to patient medical information by multiple stakeholders: physicians, nurses, allied health professionals and patients themselves. Access is rendered easily through EHRs, and an associated increase in the quality of clinical documentation is likely to follow. EHRs also enable automatic generation of clinical reminders, alerts and protocols. Furthermore, they support better management of medication, thus reducing medical errors and adverse events during clinical care.
- **Administrative efficiency**. EHRs have been called "an essential building block for healthcare's management capabilities"[21]. They can generate efficiencies in clinical practice management, such as electronic signatures, patient follow-up and telephone triage. Workflow processes benefit when multiple personnel simultaneously access the same record to support improved data intake, communication and management.
- **Cost savings**. Insufficient access to, and management of, information is responsible for many medical errors. EHRs reduce healthcare costs by lessening errors, unnecessary tests and time required to locate appropriate patient records. They also may generate associated reductions in malpractice insurance. According to one authoritative estimate, simply eliminating paper forms could save the healthcare industry up to $30 billion nationally over 10 years[22]. Most recently, experts at the Center for Information Technology Leadership (at the Partners HealthCare System in Boston) estimated a net savings of $77.8 billion nationally per year with fully implemented electronic healthcare information exchange and interoperability[23].

Respondents to the 2004 HIMSS survey reported that demonstration and research are the most effective means for establishing ROI of healthcare IT systems[24]. While relatively few academic studies demonstrating ROI on EHR implementation have been performed, the limited data available shows excellent rates of return. The areas of greatest impact appear to be reductions of drug costs and adverse drug events.

The degree of institutional benefit achieved through EHR implementation appears to depend on the reimbursement model. For example, EHRs tend to deliver lower returns in hospitals operating under fee-for-service models. Conversely, those operating under capitation plans tend to see greater value. As for institutional size, analysts have theorized that large integrated delivery systems particularly stand to gain from large EHR investments.

Erstad calls for re-engineering of work processes to realize the business potential of EHRs fully, arguing that EHR investments may have to be viewed simply as a cost of doing business. Such re-engineering should include implementation across the healthcare system, from the patient's entry into the information system during registration to the final billing process that triggers a request for services rendered. The EHR also may prove valuable as a management tool by providing information about business performance.

Common data standards could reduce the business risk inherent in EHR development. The Institute of Medicine notes, for example, that a functional EHR model is a prerequisite for vendors to develop associated management software. Better standards would enable increased interoperability between systems, tending to push down the cost of EHR systems while boosting market acceptance. Agreement on standard data elements — for example, a unique patient identifier — would result in longitudinal EHRs that supported the lifetime medical record and continuity of patient care across caregivers and institutions.

Bates et al. additionally comment that data standards would provide a financial incentive for conversion of legacy systems, a move that is crucial for processing the truly longitudinal records of citizens whose locations and healthcare providers may change annually[25]. The federal government has acknowledged that standards development should include incentives for businesses to go along, promoting the idea that savings throughout the healthcare system should be shared with those who pay for and maintain the standards.

Unfortunately, the lack of a clear EHR business case has meant that standards development itself has assumed low priority. The individuals who develop standards generally are volunteers for organizations charged with the task. Because these organizations typically receive no direct financial benefits from their efforts, as would occur in the private sector, standards may take years to materialize.

However, standards development addresses the broader, longer-term issues of controlling healthcare costs and improving quality, as well as what Bates and his co-authors call "social ROI." Investment in standards accordingly can be justified on the grounds of improving the healthcare system's medical efficacy, service efficiency and operating performance.

### The National Health Information Infrastructure

As early as 1994, Detmer cited four factors as considerable barriers to EHR development. These include:

- the lack of a national framework to address the lack of healthcare IT standards
- the disincentive to vendors to participate in standards development
- the absence of trained experts and users to disseminate good standards, and
- the lack of secure networks to transmit health information[26].

These concerns were tackled head-on by the formation of the National Healthcare Information Infrastructure (NHII.) The new federal initiative seeks to:

- advance important national goals of informing clinical practice through the use of EHRs
- interconnect clinicians to support health information exchange
- personalize care through consumer-based health records, and
- improve public health through biosurveillance.

The IT industry is explicitly identified as a stakeholder in the NHII and is called on to contribute to the infrastructure through two means. The first is leadership: "Designate internal representatives to provide strategic

leadership and coordination on issues related to NHI development and implementation."

And the second pertains to standards: "Develop and promote healthcare software and technologies that comply with national standards.[27]"

This work must be done, however, with extremely high awareness of the sensitivity of patient medical information. Confidentiality, privacy, and security as legislated through HIPAA pose significant challenges to EHR development. Waegemann has written that the original vision of a lifetime medical record has undergone a shift: "There is an understanding that only relevant information should be provided to or accessible by a particular practitioner"[28]. Kurtz has called this a "balancing act ... between ease of access for prompt medical care and ... information security to maintain security"[29].

Provision of relevant information to relevant practitioners mandates regulation of the EHR system through information security measures ranging from role-based access controls through audit trails. These requirements further accentuate the need for document data models and data exchange standards so that sensitive data may be compared and retrieved confidentially and efficiently.

## Research Methodology

Although patients' medical charts often are reviewed for quality assessment and clinical research purposes, analysis of medical records for their structure and content occurs far less frequently. A primary reason may be that these efforts typically are done by systems analysts and programmers for short-term projects. Thus they can be considered low-level knowledge work at healthcare institutions, which tend toward hierarchical organization in which executives and physicians play dominant roles.

The work done in Germany by Bludau, Hochlehnert and Wolff is an exception. These physicians identified the content and frequency of elements in 120 patient discharge letters to inform development of a data model for the Clinical Document Architecture.[30] "Surprisingly," write Lovis et al., "whereas there are numerous papers on the EPR [Electronic Patient Record] organization, the medical NLP [natural language processing] techniques and medical semantic representation, very few papers can be found about the overall structure of medical narratives themselves or the structure and typology of paragraphs used to build these narratives"[31] For example, few researchers have looked systematically at the structure of radiology reports, in which structure is understood as "not simply an ordering of the text but a strategy to meaningfully present information".[32]

For data modeling purposes, however, analysis of the current state of medical records is critically necessary. The following study was done using the methods of manual chart review and qualitative analysis using NVIVO software (QSR; Sydney, Australia; www.qsr.com.au.) Paper charts were pulled from both inpatient and outpatient medical records documenting the care of five young adult patients, who were seen continuously in a pediatric spina bifida clinic in the Northeast United States.

## Content Evolution

### Standards Organizations in Healthcare IT

Standards development in healthcare information technology dates from laboratory message exchange in the late 1960s. It originated in 1965 as the Systematic Nomenclature in Pathology, or "SNOP", a clinical vocabulary used for encoding of concepts in laboratory pathology. The somewhat curious acronym was changed to SNOMED in the 1970s to cover an expanded range of general medical and nursing concepts, rather than those exclusive to pathology. In 2004, the federal government licensed the SNOMED-Reference Thesaurus from the College of American Pathologists to support a common vocabulary standard for clinical data exchange nationwide.

Progress toward a shareable EHR also has been hampered by the current lack of national standards. Instead, interfacing solutions have been developed to govern exchange of specific data types, not overall EHR solutions. For example, DICOM (Digital Imaging and Communications in Medicine), developed by the American College of Radiology, covers image exchange; LOINC (Logical Observation Identifier Names and Codes), maintained and developed by the Indianapolis-based Regenstrief Institute, covers laboratory test results; and SNOMED covers communication of clinical concepts. Each of these standards pertains to a different type of clinical data (images, laboratory tests, diagnoses), which means that no single standard can describe the same material that is described by others.

### Health Level Seven

Health Level Seven, or HL7 (www.hl7.org), is a U.S.-based, ANSI-accredited Standards Developing Organization with international affiliates active in more than 15 nations, including Canada, China, Germany, Japan and the United Kingdom. Members are individuals, businesses and organizations within the domain of healthcare, all of which are concerned with standards for clinical and administrative data. Individual members are volunteers drawn from various healthcare sub-communities: providers, vendors, academics, consultants and government groups, with a common interest and stake in standards. Corporate members include 90% of the largest healthcare IT vendors.

HL7 emerged as an important standards developing organization in 1987 when it released its message format standards for patient registration, orders, observations and reporting. So solid is HL7's reputation that Tommy Thompson, former U.S. Secretary of Health and Human Services, in 2003 named it as one of two key health organizations to help guide development of a national EHR standard. The other entity, the Institute of Medicine, was asked to provide guidance on basic functions necessary for an EHR to promote patient safety. HL7 was to develop a set of functional requirements for an EHR system (note: *not* an EHR itself) based on that guidance.

EHR-related work within HL7 has fallen into two principal areas. After some preliminary work on an EHR standard, a special interest group (SIG) was founded in 2001, elevated to a Technical Committee several years later, and charged with work on standards towards a shareable EHR.

### The Clinical Document Architecture

In 1999, the Standardized General Markup Language (SGML) SIG created a prototype for the Clinical Document Architecture (CDA). The objective: to define common data elements of a medical record for encoding with standard SGML tags. Originally called the Patient Record Architecture (PRA), the proposal's proof-of-concept demonstration took place in early 1999. Ten healthcare vendors participated in a demonstration of SGML use for electronic medical records management at the annual HIMSS meeting in Atlanta, Georgia.[33] The PRA was renamed the Clinical Document Architecture in August of 2000.

The expressed goals of the CDA are as follows:

- Give priority to documents generated by clinicians involved in direct patient care.
- Minimize the technical barriers needed to implement the standard.
- Promote longevity of all information encoded according to this architecture.
- Promote information exchange that is independent of the underlying transfer or storage mechanism.
- Enable policy makers to control their own information requirements without extension to this specification.[34]

The CDA is considered part of HL7's "family" of standards in that its semantic content is consistent with the Reference Information Model (RIM): "a large pictorial representation of clinical data [that] identifies the life cycles of events that messages and documents convey." All HL7 standards for version 3, the most recent version at this writing, derive their content from the RIM and thus are compatible with each other.

The CDA can be viewed as a super-set of document templates hierarchically organized to prescribe the semantics and constraints on content of clinical documents. The nature of these constraints has been explicitly formalized. The intent of the CDA is to support reuse, exchange, and longevity of documents in a system-independent manner, raising the possibility of generating medical records usable throughout an individual's lifetime. In fact, when the first level of the CDA was successfully balloted by HL7's membership in October of 2000, the celebratory press release announced that this vote "brings the healthcare industry closer to the realization of an electronic medical record".[35]

However, the CDA must be understood not as an EHR in itself, but *as a data model standard enabling important attributes of clinical data to be constructed into an EHR*. These attributes are:

- accessibility
- accuracy
- comprehensiveness
- consistency
- currency
- definition
- granularity
- precision
- relevancy
- timeliness[36]

In addition, the CDA promotes a method of secure exchange of documents that have been modeled.

### Privacy Issues

The ownership, privacy and security of medical data clearly will impact development, adoption, diffusion and eventual use of EHR technologies. In fact, the National Health Information Infrastructure explicitly recognizes the social consequences of health information management by naming privacy as one of its basic components. The commitment to privacy and confidentiality actually predates Hippocrates, when it was considered essential to retain trust between patients and their healers. In modern society, medical information has a much wider audience. It is accessed for purposes of insurance, education and employment, any of which may affect an individual's life and livelihood.

Americans understandably are concerned about disclosure of such sensitive information. One recent study revealed that one out of five individuals polled believed their medical information already was being shared inappropriately.[37] These consumer attitudes have significant ramifications for the accuracy of medical data, and thus for quality improvement in healthcare. One of six Americans in the same study reports they have given inaccurate information to their healthcare provider "because they do not feel it will be kept in confidence".[38]

Omission of information is almost as big a problem as accuracy. Insufficient information has been implicated as one of many failure modes producing medical errors and adverse events in hospitals. For example, patients may withhold information regarding drugs they have been prescribed, dosages of those drugs, and known allergies to medication. In the context of HIPAA, the quality of patient medical information becomes one of potential malpractice risk when institutions provide care based on incomplete knowledge.

Goodwin and Prather report "a growing trend for patients to withhold information, and to seek care under fictitious names and erroneous social security numbers".[39] Not surprisingly, the quality of data is related to its sensitivity. Patients with sensitive information in their medical records have been found less likely to consent to disclosure.[40] Nor do clinicians necessarily trust how information may be used. Among 700 physicians likely to encounter patients with significant family histories of cancer, 29% did not want genetic test results to appear in patients' files in order to maintain confidentiality. In other words, physicians reported that they would preemptively exclude information on the patient's behalf — a clear illustration of the close connection between information sensitivity and medical record completeness.[41]

## Sensitivity

According to health data expert Thomas Rindfleisch, medical records contain "some of the most sensitive information about who and what we are."[42] Clinical narrative text may contain combinations of information (e.g., occupation and diagnosis) that make an individual's identity obvious to a knowledgeable viewer. For example, in a case study of de-identification at Duke University, "outliers (e.g., a very young pregnant girl) were still potentially identifiable by a small number of employees who knew the patient".[43] This patient's status as an outlier rendered her identifiable in that specific clinical context.

The problem of such accidental or intentional, surreptitious identification is little noted in the literature, except as a factor in other kinds of analyses. For example, Johnson and Friedman write about a natural language processing (NLP)-based investigation of expressions of patient race in discharge summaries: "In cases in which race is unknown … information from the discharge summary may be available. While the percentage of cases in which this was possible was low (1.7%), analysis of more paragraphs of the summary "e.g., 'social history') are likely to yield additional information. The 'social history' paragraph in the discharge summary often indicates country of origin, language spoken, etc."[44] In other words, race – and thus identity – may be discernable even when language about ethnicity has been sanitized.

To further complicate matters, certain types of information contained in textual narrative may embarrass even those individuals who are not positively or uniquely identified by the clinical record. Consider the following topics that one legal specialist in health information suggests for inclusion in a hospital "Release of Information" policy: abortion; adoption records; blood type and donation; infectious diseases (AIDS, HIV, venereal diseases, tuberculosis); legally or clinically incompetent adults; mental health records; organ donation; sexual assault; substance abuse records; and issues relating to status as a minor, such as pregnant minors, minors who are parents, and birth control for minors.[45]

Rindfleisch describes three principal kinds of threats to confidentiality:

- **From inside the institution**, in the form of accidental disclosures (such as conversations overheard in an elevator); insider curiosity (as in the case of the physician with legitimate reasons to access an EHR, but not a celebrity patient's record when the celebrity is being seen by another physician); or insider subornation (such as the technician who uses his access to the EHR to see the celebrity's history of substance abuse, then sells that history to the tabloid press.)
- **From outside** the institution by hackers seeking to gain medical information or disrupt systems. The celebrated case of Dutch "hacktivists" who infiltrated the EHR system at the University of Washington in 2000, simply to highlight a lack of medical information security, is a well-known example in the industry.[46] Fortunately, such incursions are relatively rare in the U.S. for the simple reason that the American healthcare industry is still almost totally reliant on paper records — possibly the only upside of paper compared with EHRs.
- **From inside, but within settings secondary** to the original information requirements, when data controls are insufficient to ensure medical information is adequately protected. Such violations may be unintentional. For example, data mining undertaken for valuable clinical research, such as detecting the frequency of adverse drug events in a particular patient population, might raise concerns of this kind.

## Data Mining

Few academic studies have explored data mining of medical narrative text. A study presented in 2003 helps explain why. Rao and Rao, representing a collaboration between medicine (University of Pittsburgh Medical Center) and industry (Siemens), performed a quality assurance study in which data was drawn from existing patient records. They cite the value of such records for clinical practice; clinical research; cost containment; quality enhancement; and identification of cohorts of patients supporting all of these objectives in situations where coding (for example, ICD-9 codes for medical billing) is unreliable.

Rao and Rao characterize narrative clinical data as unstructured, non-uniform, and non-normalized, all features that render patient records less than optimal for automated extraction purposes. The alternative approach—manual review and extraction—requires expert review, which produces data of very high quality but is prohibitively expensive. Rao and Rao conclude that "analysis of existing data is hard … At present there is no solution that allows for accuracy in the context of large numbers of patient records, by combining both structured and non-structured data."[47]

Automatic identification and extraction of clinical text has been performed successfully for reasons that include:

- retrospective studies of clinical cases to improve medical knowledge and education
- clinical decision support
- clinical practice guideline implementation
- detection of syndromes, such as epidemics or bioterrorism
- identification of patients for clinical research studies[48].

Ideally, medical researchers and businesses may seek to mine patient care databases for information on specific groups, such as men, women, Hispanics, diabetics, the very old, and so on. Results can be used to better understand each group's clinical characteristics and utilization patterns, such as comparing utilization across time and locations with that of other groups.

Knowledge discovery in text, like knowledge discovery in databases, could be used to predict the characteristics of future patients before they enter the system, and to perform epidemiological studies after patient intake has occurred. Data mining in the service of bioterrorism surveillance also capitalizes on this epidemiological potential.

In short, the data mining technology used for purposes of de-identifying records — that is, the removal of personal information so that an individual person cannot be identified — is the same technology employed to make large amounts of clinical data usable for research purposes without violating HIPAA regulations.

## HIPAA

Given the absence of prior protections safeguarding the use of medical information on paper, HIPAA was an important federal step toward ensuring secure health information transactions. The legislation both formally dictated document access and helped intensify public awareness of potential medical privacy problems. HIPAA regulations define which uses and disclosures of personal health information (PHI) must have patient authorization. Information required for treatment, payment or operations does not require such authorization.

Research, however, is subject to HIPAA. One section, the HIPAA Privacy Rule, specifies the conditions under which health information may be used or disclosed for research purposes. Research is defined as follows: "A systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge".[49]

This requires considerable changes in both clinical information management and research practice. For this reason, the American Association of Medical Colleges, representing "leading research universities, medical schools, teaching and community hospitals, as well as medical specialty and scientific societies" expressed the concerns of many researchers in writing that the Rule would "create significant obstacles to the conduct of biomedical, epidemiological, health services, and other health-related research."

A significant exception to the Privacy Rule states that health information always may be used or disclosed, for research purposes, if it has been "de-identified." In other words, health information must not be able to be traced back to a particular individual. Data that has been de-identified in this way does not require approval by institutional review boards (IRBs) at hospitals, universities and other research entities. Without IRB support, medical records research legally cannot take place. De-identification has been defined as "the practice of removing identifiers, while providing means for re-identifying individual patients or subjects if required." Researchers have two principal methods of dealing with data that must be de-identified. Under one method, according to researcher Ross Anderson, data is processed to remove identifiers, then released for "arbitrary processing by untrusted programs." Under the second method, data resides in a trusted system, permitting the posing of only a very restricted set of queries. Detailed knowledge of the application is necessary for effective use of de-identified data using either method.

Anderson reminds us of the distinction between de-identifying two key data types. Statistical data, such as the number of operations a patient has undergone, typically is compiled from current records and gives only a snapshot into an individual's life. Conversely, a database linking clinical encounters in one individual's life is "effectively impossible"[50] to de-identify because the combination of data is frequently enough to identify the patient. Sweeney strongly asserts that "de-identifying data provides no guarantee of anonymity. " For example, in a sample of medical data in which two elements were

suppressed — name and Social Security Number — 69% of the patients described were rendered completely identifiable by use of a publicly available voter list.[51]

To further complicate the problem, individuals other than the patient also must have their personal health information removed from the medical record for de-identification to take place. This includes not only members of the patient's family who may be legitimately part of the medical record, but the healthcare providers involved in the patient's care. For example, the University of Pittsburgh's "De-ID" project removes all names present in medical record text, including the names of physicians and nurses who treated him or her. This is done because the removal of all names, regardless of whose names they are, relieves concern that a patient or family member's name could be confused with (or be identical to) that of a physician or other member of the healthcare team.

Melissa Saul, a De-ID developer, adds that in cases of quality assurance and review involving a physician's group practice, a business case can be made specifically for physician de-identification. In her view, the reputation of a physician in a group practice might be compromised if identifiable quality control data was released to the entire medical practice group.[54]

Unfortunately, de-identification — though "tedious, time-consuming, and expensive"[55] — is the only post-HIPAA means by which personal health information can be made available for research. This results in tensions between competing social goods: the good of scientific advancement, and the good of patient confidentiality. Many researchers fear the outcome may be reduced availability of clinical data, and thus less ability to serve the public, now and in the future.

### De-identification and HIPAA

Individually identifiable health information is defined under HIPAA to be "Information that identifies an individual; or with respect to which there is a reasonable basis to believe that the information can be used to identify the individual." The Privacy Rule stipulates 18 specific data elements that must be removed for de-identification to take place. These elements are:

- Names
- Elements of dates, except years, directly related to an individual
- Telephone and fax numbers
- Geographic subdivisions
- Electronic mail addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- URLs and IP address numbers
- Biometric identifiers
- Full-face images
- "Any other unique identifying number, characteristic or code. "

The close resemblance between the HIPAA data elements and the basic data elements required by healthcare information systems is documented in a Cisco report. She surveyed 19 healthcare institutions using document imaging (as opposed to text processing) systems. Fifty-one percent of respondents routinely captured between eight and 12 indexing values per document (range for all respondents was between three and 18). Most of these values were either manually keyed, or downloaded from the healthcare system's Master Patient Index (MPI), and the indexing fields most frequently assigned were patient name, document type, and patient MPI unit number.[56]

Table 1 shows the data typically captured, or entered, to accompany these document images. This gives a snapshot of the kind of data in typical clinical documents. Fourteen of the top 20 data elements are identified as requiring de-identification under HIPAA, and would need to be removed for de-identification to take place. The most frequently occurring data elements are also the ones that would have to be removed to make the documents available for research purposes.

Several researchers have already published results of experiments in automated deidentification.[57, 58, 59] One work group de-identified proper names in a free-text surgical pathology database: "Proper names in the free-text database were identified either from available lists of persons, places, and institutions, or by their proximity

to keywords, such as Dr. or hospital."[lx] Unfortunately, as Taira and colleagues have pointed out: "This de-identification process is tedious if performed manually, and is known to be quite faulty in direct search and replace strategies"[lxi]. Taira et al. have aptly summarized the state of de-identification in the domain of patient records as "a lax standard" given that the typical method is to perform a global search-and-replace action using the patient's identifying information as the searched-for text.

Only a few researchers working in natural language processing have begun work in this domain, and most involves relational data instead of narrative text. According to medical informatics and records expert William Stead of Vanderbilt University, the present state of the art is "a significant barrier" to full utilization of data collected in routine clinical practice.

The De-ID engine developed at the University of Pittsburgh — and now produced commercially — illustrates how such an automatic de-identifier can work with text[lxii]. De-ID replaces any text it suspects to be a personal name, date, age, place (e.g., city) or address (e.g., street address) with a series of asterisks and a label denoting what was replaced. For example, the patient's name, Ms. Smith, becomes

Ms. **NAME<AAA>

Ages that are stated in the document also are replaced by a statement of the individual's age in decades, so that a female person described in the original record as 65 years old would be rendered by De-ID:

**AGE<in 60s>-year-old woman

This can, and does, occasionally result in nonsensical processed narrative, particularly in situations involving an eponymous medical test, instrument, or procedure. For example, "Parkinson's Disease" would be translated by De-ID as ***NAME***Disease. So, too, are medical acronyms and abbreviations, such as "CVA" (cerebrovascular aneurysm, costovertebral artery, etc.) and "S/P" (status post); initially understood by De-ID to be personal names. However, its developers report that machine learning can train De-ID to perform with greater refinement.

The paucity of research in this field, in the context of the HIPAA guidelines, underscores a distinction between anonymization and desensitization. Anonymized means that the named person is rendered unrecognizable by a change in the text. For that reason, De-ID is a computerized example of anonymization. To make material desensitized, however, would mean hiding information that has potentially embarrassing or harmful implications when viewed by unintended audiences. Although the guidelines provide a clear path to anonymization through the removal of obvious elements from patient records, they provide little guidance with respect to suitable procedures for desensitizing clinical narrative within those records. In fact, the HIPAA legislation steers clear of specifying any method of de-identification of any clinical information, narrative or not, in any medium — paper or electronic.

The CDA data modeling standard potentially could alleviate the burden of HIPAA-compliant data mining for research and business purposes, while continuing to protect the patient's right to privacy and confidentiality of medical information. To enable such a future, we must improve our understanding of the content of typical paper-based clinical documents within current medical records.

The study section of this report provides an illustration of the typical, paper-based medical record found in a small urban hospital. The aim is to demonstrate what kinds of HIPAA-sensitive data are presently retained, and how the structure of clinical documents might assist in future de-identification.

## The Study

This study reports results of an analysis of paper-based clinical documents that describe the medical history and clinical encounters among a small group of patients suffering from complex, chronic illnesses. The goal was to establish the characteristics and frequency with which 18 HIPAA-sensitive data elements currently occur in patients' paper medical records at a prototypical hospital.

Chronic illness has been defined as "a condition that interferes with daily functioning for more than three months in a year, causes hospitalization for more than one month a year, or (at the time of diagnosis) is likely to do so."[60] Such care presents a significant challenge to information management, particularly given the post-HIPAA environment. Patients often require the cooperation of multiple specialist caregivers, each of whom periodically may need access to any available patient information. In addition, care may be relocated from the hospital to the home, a long-term care facility or similar site — or vice-versa. The patient also may require more support and treatment services in one setting or the other.

The combination of chronic illness with multiple medical and allied health specialties — occurring over a longer life trajectory — generates significant quantitative volume and qualitative complexity of medical documentation. This reality creates a rich field for research into the content and sensitive nature of medical data now locked in paper records. In fact, the clinic coordinator has an almost technological function: to "rapidly and effectively communicate patient information among the caregivers at the clinic and in the patient's local community ... a large responsibility, [requiring] a sizeable effort".[64]

Mid-20th century developments in medical care have created what Dosa calls a "pioneer survivor" group: people who have grown into adulthood with a chronic health condition so severe that in previous generations they would not have survived.[65] Examples of such chronic conditions include cystic fibrosis, diabetes mellitus, juvenile rheumatoid arthritis, sickle cell anemia, cerebral palsy, epilepsy, childhood cancers, congenital heart disease, and spina bifida. According to Dosa, "The need for patients, families and caregivers working in chronic illness situations to understand their own medical information is particularly acute."

The medical records of persons with spina bifida are particularly interesting for study of medical care documentation and information handling. Patients with this diagnosis must navigate a diverse array of healthcare specialties addressing its principal sequelae: hydrocephalus, varying levels of paralysis and urinary and bowel incontinence, and other serious conditions. Because medical advances are increasing survival rates among children born with spina bifida dramatically, complexity of their records likewise is rising almost geometrically. Four of the five patients in this study have diagnoses of spina bifida; the fifth is diagnosed with quadriplegia and cared for by the same medical team.

### Overview of case study

#### Site specifics

The clinical documents analyzed here were obtained from a small academic medical center hospital allied with a medical school in the Northeast United States. This 329-bed hospital incorporates a pediatric specialist clinic primarily serving spina bifida patients, as well as those with other similarly disabling and incurable conditions causing paraplegia and quadriplegia. Clinic patients range from infants to adults. The study reviewed records for both outpatient care and, when applicable, for inpatient care. Records are maintained by the larger hospital in which the specialist clinic is located, but include clinical documents generated by these patients' specialists.

The five patients are anonymized here as Patients A, B, C, D, and E. Each had been seen at the specialty clinic as a hospital outpatient, although not all had been inpatients.

#### Data specifics

*Document types*

There was considerable variation in the size and scope of each patient's record, ranging from admission forms to handwritten notes. To facilitate analysis, documents were categorized in one of two ways. "Standard" documents were official forms, which usually were preprinted. "Nonstandard" documents could be any other kind of material present in the record. These varied considerably: original letters, copies of forms and letters, fax cover sheets and even a form used as scratch paper

to record anatomical measurements. Other Nonstandard materials included numerous clinical images and graphs from ultrasonic bladder scan procedures and EKGs. Table 2 shows the distribution and scope of inpatient and outpatient Standard documentation in this sample. A summary of the proportion of Nonstandard and Standard documents found in each patient's complete medical record of appears in the appendix in Table 3.

Only 15 unique types of Standard documents could be classified as produced by healthcare providers outside the hospital. Six emanated from an outlying hospital. These were primarily registration and authorization forms, but included data related to emergency room admissions. One related to pre-hospital care by a rural ambulance service. Two were produced by the State Department of Social Services, another came from the County Health Department, and one from a second County Health Department's public health unit regarding nursing care. Two different community services programs were represented by three different documents. Finally, a standard Department of Human Health and Services pediatric growth chart was included in one patient's file.

Within the Standard category of documents, there were 218 unique forms present in the five patients' combined records. The smallest number of unique forms for one patient was 17 (for someone with no inpatient records) and the largest number was 82. Each record surveyed in this study contained an average of 55 unique Standard documents (with "unique" understood to mean "differing," not "nonstandard".)

Table 4 presents the complete list of Standard document types and the frequency with which they appeared in the entire sample of 1,010 Standard documents reviewed. In the entire sample, 42% consisted of the 10 most frequently appearing types. Prototypes were collected for all of these. The other 208 types were represented in the entire sample of documents in proportions ranging from 1.29% to 0.1%.

The 12 document types that appeared with the top 10 frequencies, together with the proportion of times they appeared in the records of each patient, is detailed in Table 5.

Very little unstructured narrative text was found in this sample. Of the 136 prototype documents examined, only 10 (7.3 %) consisted of free-text narrative. The remaining 126 (93.7%) were printed forms with checkboxes and relatively standardized fielded data.

*HIPAA data elements*

To determine the frequency of occurrence of the 18 data elements specified for de-identification of personal health information (PHI) under HIPAA, the 136 prototype Standard documents were reviewed manually. Qualitative analysis was done using NVIVO. Table 6 lists the 18 elements themselves and the corresponding data elements coded in this analysis. The same documents were reviewed for the presence and location of HIPAA sensitive data elements. Table 7 gives the frequency with which specific elements appeared.

The average document reviewed had a total of four elements, while individual documents ranged from 35 to none. Only 12 prototype documents had no HIPAA elements at all. Table 8 gives a detailed breakdown of each HIPAA element and the number of instances and prototype documents in which it appeared.

Finally, Table 9 lists the five most "sensitive" clinical documents present in the 136 prototypes. Unsurprisingly, registration records, which are necessarily rich with personal information, headed the list.

## Results

The typical patient reviewed in this study was represented by 55 unique standardized, form-based documents in his or her inpatient record, and by 196 in his or her outpatient record. An average of 5% of all documents in each patient's record consisted of data appearing in nontraditional, non-official, and in some cases non-paper format. On average, 93% of the documents reviewed were standardized official forms.

Only 15% of the 218 official types of documents were produced outside the subject hospital. The 10 most frequently appearing unique document types accounted for only 42% of the sample; more than half the types reviewed appeared in extremely low frequencies, from 1.29 to .1% of the entire sample. Each patient medical record contained an average of 55 unique documents. Among the documents reviewed, 94% consisted of printed forms in which data reporting relied on handwriting and checked boxes. The field names used in these forms were relatively standardized. Of the 18

data elements specified as "sensitive" and requiring de-identification before release for research, Name was the most frequently occurring element, followed by Medical Record Number. Name occurred in 89% of documents; another 4% contained a portion of the name, such as the first name, last name, or maiden name only. Ten percent of the survey documents contained a signature. As discussed earlier, de-identification under HIPAA relates to individuals other than patients alone. For example, Element #1, "Names", translates to removal not only of the patient's name from the medical record, but of the names of his or her family members. The representation of other individuals in this data would have to be taken into consideration during a de-identification procedure. Seven percent of documents contained the name of a patient's family member, 1% the name of an emergency contact (typically a family member), and 11% contained the signature of that family member. Although the patient's Social Security Number was found in only 4% of the documents, Medical Record Number, in many healthcare systems synonymous with SSN, was found in 82% of documents.

### CDA for Data Mining

This data clearly illustrates the nature of representative medical record documentation now. There was extreme variability in the number of standardized official documents found in this sample, and an average of 55 unique documents per patient. When all documents exchanged within a healthcare information system conform to one data model with known and standardized elements, de-identification theoretically will be easier to achieve.

The CDA can assist by providing a standardized structure for clinical document content, including specifying the section headings, or labels, for individual sections and subsections. Used in combination with knowledge about the frequency and location of HIPAA-sensitive data elements, it would be possible to automate search and pseudonymous replacement not only of data elements of text, but of sensitive sections of documents containing sensitive text. (Pseudonymization differs from anonymization in that the former replaces one name with another, while anonymization translates them into a placeholder such as "Name.") For example, a patient with the surname "Armstrong" would be pseudonymized as "Doe." Given conversion of legacy systems, it theoretically is possible to identify, "scrub", and fragment a patient's sensitive data elements and document components that recur over long time spans.

In fact, the considerable repetition of document types and data elements (in 42% of documents reviewed here) comprises the paper-based longitudinal medical record. Understanding the nature and type of data elements present in these historical documents would permit clinically meaningful comparisons and aggregation of the clinical data, while reducing the resource-intensive waste involved in repeated collection of the same elements. XML style sheets and CDA templates can format specific sections of de-identified documents for display and use in clinical information retrieval. Most importantly, these documents can be customized for exchange within a healthcare organization or between separate organizations.

Finally, structuring documents according to the CDA permits deployment of information tailored to the needs of individuals requesting it – whether that person is the primary care practitioner, specialist physician, nurse, physical therapist, dietitian, or patient. Documents that are intelligently structured, with expertise encoded about the uses to which data will be deployed, have the potential to make clinical information exponentially more useful – for individuals, the general public and the healthcare system itself.

### Recommendations

Most clinical documents are restricted from release under HIPAA for research purposes unless subjected to expensive and intensive de-identification. Routinely collected information — such as that taken at hospital registration and intake — should be analyzed to help develop a data model for electronic information exchange under national healthcare information standards. Considerable efficiencies simultaneously could be realized through internal analysis of repetition and overlap between these unique documents.

## Conclusion

In medical records, high-quality data elements are characterized by adherence to accepted standards: accessibility, accuracy, comprehensiveness, consistency, currency, granularity and precision. In addition, clinical data must be structured for automated clinical reminders and decision support tools to save lives and maximize financial resources. Structuring requires standardization for data modeling and data exchange.

The lack of standards for patient medical information constitutes a significant barrier to implementation of EHRs. The absence of standards results in decreased interoperability among and between health information systems, as well as lowered quality, accountability, and overall integrity of existing data. As a result, creating a true longitudinal EHR throughout the patient's lifespan is virtually impossible. The future may consist of EHR data silos unless industry stakeholders agree on development and maintenance of clinical data standards.

Structuring clinical documentation to enable institutions to share clinical data offers a significant ancillary benefit: greatly protecting data privacy and security. By expanding knowledge about the location, content and frequency of data elements in personal health information, chances that sensitive information will be released inappropriately are greatly reduced. The result could be far stronger protection for individuals, their families, and the care providers and institutions earnestly seeking to improve their patients' health.

# Appendix

## Useful websites

Health Level Seven

http://www.hl7.org

The Data Privacy Lab at Carnegie Mellon University

http://lab.privacy.cs.cmu.edu/people/sweeney/

**Table 1**

Indexing Fields Captured or Entered in Typical Document Imaging Systems

| Indexing Value | Respondents |
|---|---|
| Patient name | 33 (94%) |
| Document type | 32 (91) |
| Patient MPI unit number | 31 (89) |
| Date of service | 29 (83) |
| Date of encounter/admission | 27 (77) |
| Patient date of birth | 25 (71) |
| Date of discharge | 21 (60) |
| Patient account number | 20 (57) |
| Patient social security number | 18 (51) |
| Patient MPI encounter number | 18 (51) |
| Caregiver (e.g., admitting physician) | 13 (37) |
| Insurance/health plan | 11 (31) |
| Document page number | 9 (26) |
| Document date | 7 (20) |
| Document ID number | 6 (17) |
| Chief problem/Primary diagnosis | 5 (14) |
| Type of encounter | 4 (11) |
| Patient gender | 3 (9) |
| Patient type | 2 (6) |
| Secondary problem/diagnosis | 2 (6) |
| Time of service | 2 (6) |
| Patient race | 1 (3) |
| Patient financial class | 1 (3) |
| User access | 1 (3) |
| Document name | 1 (3) |
| Bar code form number | 1 (3) |

**Table 2**

Inpatient and Outpatient Standard Documentation in 5 Patients' Medical Records

| Patient | Inpatient | Outpatient |
|---|---|---|
| A | n/a | 30 |
| B | 109 | 207 |
| C | 52 | 336 |
| D | 114 | 291 |
| E | n/a | 120 |

**Table 3**

Nonstandard vs. Standard Documents in 5 Patients' Medical Records

| Patient | Nonstandard | (%) | Standard | (%) | Total |
|---|---|---|---|---|---|
| A | 3 | 9% | 30 | 91% | 33 |
| B | 30 | 9% | 315 | 91% | 345 |
| C | 35 | 8% | 384 | 92% | 419 |
| D | 13 | 3% | 405 | 97% | 418 |
| E | 6 | 5% | 120 | 95% | 126 |

**Table 4**

Standard Documents found in the Sample

| Standard Document Type | Total | Proportion of all Standard documents (N=218) |
|---|---|---|
| Registration Record | 120 | 11.88% |
| Department of Radiology Consultation Report | 36 | 3.56% |
| Letter of Justification: Seating and Positioning Program | 30 | 2.97% |
| Medication Administration Record | 27 | 2.67% |
| Physicians Orders | 27 | 2.67% |
| Progress Notes | 27 | 2.67% |
| Outpatient Visit Note | 23 | 2.28% |
| Diagnostic X-ray report | 20 | 1.98% |
| Emergency Department Record | 20 | 1.98% |
| 24-hr flow sheet | 17 | 1.68% |
| Consent to Diagnostic and Medical Treatments; Release of Information, Medicare Benefits, and Financial Agreement | 16 | 1.58% |
| Emergency Nursing Record | 16 | 1.58% |
| Patient Discharge Instructions | 16 | 1.58% |
| Cumulative Summary | 15 | 1.49% |

**Table 5**

Twelve Most Frequently Occurring Standard Document Types

| | | # in patient records | | | | |
|---|---|---|---|---|---|---|
| **Standard document** | **% of sample (N= 1010)** | **A** | **B** | **C** | **D** | **E** |
| Registration Record | 11.88% | 6 | 39 | 49 | 14 | 14 |
| Department of Radiology Consultation Report | 3.56% | 2 | 3 | 6 | 24 | 1 |
| Letter of Justification: Seating and Positioning Program | 2.97% | 2 | 11 | 12 | 2 | 3 |
| Medication Administration Record | 2.67% | 0 | 11 | 1 | 15 | 0 |
| Physicians Orders | 2.67% | 0 | 12 | 4 | 12 | 0 |
| Progress Notes | 2.67% | 0 | 12 | 9 | 6 | 2 |
| Outpatient Visit Note | 2.28% | 1 | 6 | 10 | 3 | 3 |
| Diagnostic X-ray report | 1.98% | 0 | 5 | 15 | 0 | 0 |
| Emergency Department Record | 1.98% | 0 | 7 | 8 | 2 | 3 |
| 24-hr flow sheet | 1.68% | 0 | 17 | 0 | 0 | 0 |
| Consent to Diagnostic and Medical Treatments; Release of Information, Medicare Benefits, and Financial Agreement | 1.58% | 0 | 2 | 3 | 5 | 4 |
| Emergency Nursing Record | 1.58% | 0 | 7 | 7 | 0 | 2 |

**Table 6**

18 HIPAA-specified elements and variants identified in this study

| **Elements** | **Includes** |
|---|---|
| Names | Patient name |
| | Patient maiden, first or last name only |
| | Family member name |
| | Emergency contact name |
| | Patient signature |
| | Family member signature |
| | Other individual signature |
| | Multiple identity signature |
| Elements of dates, except years, directly related to an individual | Patient date of birth |
| | Patient family member date of birth |
| | Telephone and fax numbers |
| | Patient telephone number |
| | Family member telephone number |
| | Emergency contact telephone number |
| | Guarantor telephone number |
| | Other individual telephone number |
| Geographic subdivisions | Patient address |
| | Family member address |
| Electronic mail addresses | Electronic mail addresses |

**Table 6 (continued)** 18 HIPAA-specified elements and variants identified in this study

| Elements | Includes |
|---|---|
| Social Security numbers | Patient Social Security #<br>Policy holder Social Security #<br>Medical record # |
| Medical record numbers | Medical record # |
| Health plan beneficiary numbers | Health plan beneficiary # |
| Account numbers | Account numbers |
| Certificate/license numbers | Certificate/license numbers |
| Vehicle identifiers and serial numbers, including license plate numbers | Vehicle identifiers and serial numbers, including license plate numbers |
| Device identifiers and serial numbers | Device identifiers and serial numbers |
| URLs and IP address numbers | URLs and IP address numbers |
| Biometric identifiers | Biometric identifiers |
| Full-face images | Full-face images |
| "Any other unique identifying number, characteristic or code" | "Any other unique identifying number, characteristic or code"<br>Patient Record #<br>Medicaid #<br>Medicare #<br>Health Plan Beneficiary # |

**Table 7**

Instances of HIPAA Elements Found in All Prototype Documents (n=136)

| Element | # |
|---|---|
| Names | 260 |
| Elements of dates, except years, directly related to an individual | 47 |
| Telephone and fax numbers | 21 |
| Geographic subdivision | 66 |
| Electronic mail addresses | 0 |
| Social Security numbers | 14 |
| Medical record numbers | 112 |
| Health plan beneficiary numbers | 2 |
| Account numbers | 9 |
| Certificate/license numbers | 0 |
| Vehicle identifiers and serial numbers, including license plate numbers | 0 |
| Device identifiers and serial numbers | 0 |
| URLs and IP address numbers | 0 |
| Biometric identifiers | 0 |
| Full-face images | 0 |
| "Any other unique identifying number, characteristic or code" | 16 |

**Table 8**

HIPAA Elements Present in Prototype Documents

| Element | Instances | Documents | Proportion of documents |
|---|---|---|---|
| Names | | | |
| Patient name | 154 | 121 | 89% |
| Patient maiden name only | 1 | 1 | 1% |
| Patient first name only | 27 | 3 | 2% |
| Patient last name only | 1 | 1 | 1% |
| Family member | 14 | 10 | 7% |
| Emergency contact name | 2 | 1 | 1% |
| Patient signature | 17 | 14 | 10% |
| Family member signature | 15 | 15 | 11% |
| *Subtotal* | *231* | *166* | |
| Geographic Subdivisions | | | |
| Patient | 63 | 61 | 45% |
| Family member | 3 | 2 | 1% |
| SUBTOTAL | 66 | 63 | |
| DATES | | | |
| Patient DOB | 46 | 43 | 32% |
| Patient family member DOB | 1 | 1 | 1% |
| *Subtotal* | *47* | *44* | |
| Other Identifying Numbers | | | |
| Social security # of policy holder | 5 | 2 | 1% |
| Social security # of patient | 9 | 5 | 4% |
| Medical record # | 96 | 112 | 82% |
| Health plan beneficiary number | 2 | 1 | 1% |
| Account # | 9 | 9 | 7% |
| Other #, characteristic code | 1 | 1 | 1% |
| Medicare # | 1 | 1 | 1% |
| Medicaid # | 1 | 1 | 1% |
| Patient Record # | 1 | 1 | 1% |
| *Subtotal* | *125* | *133* | |
| Communications Media Numbers | | | |
| Telephone numbers | | | |
| Patient telephone number | 10 | 9 | 7% |
| Family member telephone number | 4 | 4 | 3% |
| Guarantor telephone number | 1 | 1 | 1% |
| Emergency contact telephone number | 2 | 1 | 1% |
| Other individual telephone number | 2 | 1 | 1% |
| Unknown phone number | 2 | 1 | 1% |
| *Subtotal* | *21* | *17* | |

**Table 9**

Documents containing the greatest number of HIPAA elements

| Document | Elements (per document) |
|---|---|
| Hospital Outpatient Registration Record | 35 |
| Department of Neurosurgery | 25 |
| Registration Record | 16 |
| Emergency Department Admission Record | 16 |
| Patient's Personal Property Record | 13 |

**Acknowledgments**

The author expresses her sincere appreciation to Nienke Dosa, MD, Thomas Welch, MD, and the staff of the Department of Pediatrics at Upstate Medical University for facilitating this research.

## Endnotes


[1] Spiegel, A.D, & Springer, C. R. (1997). Babylonian medicine, managed care and Codex Hammurabi, circa 1700 B.C. Journal of Community Health, 22(1), 69-89.

[2] Frisse, M. (1992). The health of the computer-based patient record. Academic Medicine, 67(7), 441-443.

[3] Morrissey, J. (2001). Vendors say they're ready to deliver. Modern Healthcare, 31(46), 28-30.

[4] National Committee on Vital and Health Statistics [NCVHS]. (2001, November 15). Information for health: A strategy for building the National Health Information Infrastructure. Washington, DC: U.S. Department of Health and Human Services. Available online: http://aspe.hhs.gov/sp/nhii/Documents/NHIIReport2001/default.htm [Date accessed: January 31, 2005].

[5] Bates, D.W., Ebell, M., Gotlieb, E., Zapp, J., & Mullins, H.C. (2003). A proposal for electronic medical records in U.S. primary care. Journal of the American Medical Informatics Association, 10(1), pp. 1-10.

[6] Carpenter, D. (2002). The paperless chase. Hospitals & Health Networks, 76(1), 47.

[7] Valdes I, Kibbe DC, Tolleson G, Kunik ME, Petersen LA. (2004). Barriers to proliferation of electronic medical records. Informatics in Primary Care, 12(1), pp. 3-9.

[8] HIMSS EHR Committee. (2003, June 30). HIMSS EHR Definitional Model v. 1.0. p. 2. Available online: http://www.medical.siemens.com/siemens/en_US/gg_hs_FBAS/files/brochures/pdf_files/EMRdefinition.pdf. [Date accessed: January 31, 2005].

[9] Ondo, K., Wagner, J., & Gale, K. The electronic medical record: Hype or reality? Journal of Healthcare Information Management, 17(4): p. 2.

[10] Bird, L.J., Goodchild, A., & Beale, T. (2000). Integrating health care information using XML-based metadata. [Unpublished manuscript]. Available online: http://citeseer.nj.nec.com/bird00integrating.html. [Date accessed: Jan. 31, 2005].

[11] Cisco, S.L. (1996). Document imaging in the United States: A survey of several hundred hospital installations. (p. 8). Newton, MA: Medical Records Institute.

[12] Institute of Medicine (2003). Key Capabilities of an EHR System. P. 2.

[13] HIMSS, Definitional Model (2003).

[14] NCVHS, op. cit.

[15] Fowles, ,J.B., Kind, A.C., Craft, C., Kind, E.A., Mandel, J.L., & Adlis, S. (2004). Patients' interest in reading their medical record: relation with clinical and sociodemographic characteristics and patients' approach to health care. Archives of Internal Medicine, 164(7):793-800.

[16] Federal focus on healthcare information technology. (2004, July). Vantage Point. Available online: http://www.himss.org/content/files/vantagepoint/vantagepoint_072004d.htm. [Date accessed: January 31, 2005].

[17] NCVHS, p. 13.

[18] Bates et al., op. cit.

[19] Vogel, LH. (2003). Finding value from IT investments: exploring the elusive ROI in healthcare. Journal of Healthcare Information Management, 17(4):20-8.

[20] Lang, R.D. ROI and IT: strategic alignment and selection objectivity. Journal of Healthcare Information Management, 17(4):2-3.

[21] Dettmer, D.E. (1994). Computer-based patient records: A building block for healthcare reform. Physician Executive, 20(1), p. 16.

[22] Hester, R.D. (2003). Physician medical records and the Health Insurance Portability and Accountability Act. Journal of Health and Social Policy, 18(1), pp. 1-14.

[23] Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D.W., & Middleton, B. (2005). The value of health care information exchange and interoperability. Health Affairs,24(1). Available online: http://content.healthaffairs.org/cgi/reprint/hlthaff.w5.10v1 [Date accessed: Jan. 31, 2005].

[24] Federal focus on healthcare information technology, op. cit.

[25] Bates et al., op. cit.

[26] Detmer, DE. (1994). Computer-based patient records: a building block for health care reform. Physician Executive, 20(1), pp. 16-9.

[27] NCVHS, Information for health, op. cit.

[28] Waegemann,C.P. (2002), p. 64.

[29] Kurtz, G. (2003). EMR confidentiality and information security. Journal of Healthcare Information Management, 17(3), pp. 41-48.

[30] Bludau, H-B., Hochlehnert, A.J., & Wolff, A. CDA XML-based discharge letters, University of Heidelberg. Paper presented at the HL7 International CDA Conference, October 7-9, 2002, Berlin, Germany. Available online: http://www.hl7.de/cda2002/absbiopres/bludau.html [Date accessed: Jan. 31, 2005].

[31] Lovis, C., Baud, R., Revillard, C., Pult, L., Borst, F., & Geissbuhler, A. (2001). Paragraph-oriented structure for narratives in medical documentation. Medinfo, p 639.

[32] Rooksby, J., & Kay, S. (2001). Clinical narrative and clinical organisation: Properties of radiology reports. Medinfo, pp. 683.

[33] Lincoln, T., Spinosa, J., Boyer, S., & Alschuler, L. (1999). HL7-XML progress report. In XML Europe '99 Conference Proceedings (pp. 733-6). Alexandria, VA: Graphic Communications Associates.

[34] Dolin, R. H., Alschuler, L., Beebe, C., Biron, P. V., Boyer, S. L., Essin, D., Kimber, E., Lincoln, T., & Mattison, J. E. (2001). The HL7 Clinical Document Architecture. Journal of the American Medical Informatics Association, (8)6, pp. 553-554.

[35] Health Level Seven. (2000). HL7 to release first XML-based standard for healthcare. [Press Release]. Ann Arbor, MI: Health Level Seven.

[36] NCVHS, Report to the Secretary, p. 35. Hester RD. Physician medical records and the Health Insurance Portability and Accountability Act. Journal of Health and Social Policy, 18(1):1-14, 2003.

[38] Hester, op. cit.

[39] Goodwin, L.K., & Prather, J.C. Protecting patient privacy in clinical data mining. Journal of Healthcare Information Management, 16(4), p. 62.

[40] Merz, J.F., Spina, B.J., Sankar, P. Patient consent for release of sensitive information from their medical records: an exploratory study. Journal of the Behavioral Sciences & the Law, 17:445-454, 1999.

[41] Wasserman, L.M., Jones, O.W., Trombold, J.S., & Sadler, G.R. (2000). Attitudes of physicians regarding receiving and storing patients' genetic testing results for cancer susceptibility. Journal of Community Health, 25(4), 305-313.

[42] Rindfleisch, op. cit., p. 94.

[43] Goodwin & Prather, op. cit., p. 67.

[44] Johnson, S.B., & Friedman, C.P. (1996). Integrating data from natural language processing into a clinical information system.Proceedings of the AMIA Annual Symposium, 537-41.

[45] Carman, D.M. Balancing patient confidentiality and release of information. Bulletin of the American Society of Information Science, 23(3).

[46] Hacker accesses patient records. (2000, December 9). Washington Post, p. E1.

[47] Rao, R.H., & Rao, R.B. (2003). Quality assurance through comprehensive extraction from existing (nonstructured) patient records. Paper presented at the annual meeting of the Health Information Management Systems Society. Available online: http://www.himss.org/content/files/proceedings/2003/Sessions/session106_slides.pdf [Date accessed: 8 June 2005].

[48] Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., & Buchanan, B.G. (2001). Evaluation of negation phrases in narrative clinical reports. Proceedings of the AMIA Annual Symposium, 105-109.

[49] HIPAA, Privacy Rule Final Modifications, 2002.

[50] American Association of Medical Colleges. (2002, April 11). AAMC comment letter on privacy NPRM. Available online: http://www.

aamc.org/advocacy/library/hipaa/corres/2002/041102.htm. [Date accessed: June 8, 2005].

[51] Anderson, R. The DeCODE Proposal for an Icelandic Health Database. 1998, October. Available online: http://www.cl.cam.ac.uk/users/rja14/iceland/iceland.html.

[52] Anderson, 1998, op. cit.

[53] Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, p. 2.

[54] Melissa Saul, personal communication, October 2004

[55] Goodwin & Prather, op. cit., p. 66.

[56] Cisco, op. cit., p. 5.

[57] Goodwin & Prather, op. cit.

[58] Taira, R.K., Bui, A.A., & Kangarloo, H. (2002). Identification of patient name references within medical documents using semantic selectional restrictions. Proceedings of AMIA Annual Symposium, 757-61.

[59] Thomas, S.M., Mamlin, B., Schadow, G., & McDonald, C. (2002). A successful technique for removing names in pathology reports using an augmented search and replace method. Proceedings of AMIA Annual Symposium, 777-81.

[60] Miller, R.E., Boitnott, J.K., & Moore, G.W. (2001). Web-based free-text query system for surgical pathology reports with automatic case de-identification. Archives of Pathology and Laboratory Medicine, 125:8.

[61] Taira et al., p. 757.

[62] Gupta, D., Saul, M., & Gilbertson, J. (2002). Evaluation of a de-identification software engine: Progress towards sharing clinical documents and pathology reports. Proceedings of the AMIA Annual Symposium.

[63] Perrin, J.M. (1985). Introduction. In N. Hobbs & J.M. Perrin (Eds.), Issues in the care of children with chronic illnesses (pp. 2-10). San Francisco: Jossey-Bass.

[64] Kaufman, B.A., Terbrock, A., Winters, N., Ito, J., Klosterman,. A., & Park, T.S. (1993). Disbanding a small multidisciplinary clinic: Effects on the health care of myelomeningocele patients. Pediatric Neurosurgery, 21, p. 40.

[65] Dosa, N. (2002). Pioneer-survivors: Insights on childhood resilience by adults with spina bifida. (Unpublished manuscript). Syracuse, NY: University Hospital.

## About the Author

Catherine Arnott Smith is an Assistant Professor in the School of Information Studies at Syracuse University. She holds masters degrees in library science and American history (University of Michigan, 1992) and information science (University of Pittsburgh, 2000). Between 1997 and 2002 Dr. Arnott Smith held a predoctoral research fellowship at the Center for Biomedical Informatics at the University of Pittsburgh, from which she received her PhD.

In 2002, Dr. Arnott Smith became the first Donald A.B. Lindberg Research Fellow. The research fellowship is endowed by the Medical Library Association and named in honor of the Director of the National Library of Medicine. This award supports Dr. Arnott Smith's ongoing Ten Thousand Questions Project, which analyzes the content and terminology of consumer questions on free, public Web-based bulletin boards.

She is Chair and Chair-Elect, respectively, of the Medical Informatics Sections of the Medical Library Association and the American Society for Information Science and Technology. She has published in the Journal of the American Society for Information Science and Technology, the Journal of the Medical Library Association, and the Proceedings of the American Medical Informatics Association, and contributed a chapter to Consumer Health Informatics, the first textbook in this field, published in 2005 by Springer-Verlag.

At Syracuse, Dr. Arnott Smith teaches courses in library systems and processes, consumer health and medical vocabularies to graduate students in library science, information management, information science, and technology and public policy. Dr. Arnott Smith's current research interests are (1) Data modeling of clinical information; (2) consumer health vocabulary; (3) controlled medical terminologies, particularly in unexplored domains such as alternative/complementary medicine; and (4) the interface between consumers and their electronic medical records.

Printed in the United States of America
07-05


HCM07-00