Information retrieval in medicine:
The electronic medical record as a new domain

Catherine Arnott Smith, MA, MILS, MSIS, PhD
Assistant Professor
School of Library and Information Studies
University of Wisconsin-Madison
Room 4217 Helen C. White Hall
600 N. Park Street, Madison, WI 53706

Abstract

"The medical record is a material form of public memory," Berg (1996) writes, "a structured distributing and collecting device, where all tasks concerning a patient's trajectory must begin and end…" [Italics original; p. 510]. Structured distributing and collecting devices are the natural interest of information science. Unfortunately, of the 130 articles published about medicine in almost 36 years of JASIST, although 70 (54%) deal with information retrieval, communication and the work processes behind them, only 2 of these articles (1.5%) have focused on the medical record.

The body of existing information retrieval work most relevant to the medical record as a base for experiment is the work called "passage retrieval" defined as "the task of identifying and extracting fragments from large, or short but heterogeneous full text documents" (Melucci, 1998, p. 44).

This paper presents a document-centered approach to the EHR as an information retrieval problem. It is clear that passage retrieval researchers working in the field of information science have seen similar values in document passages as have researchers in medical informatics. Without either literature acknowledging the other, workers in both camps have identified the same potential in document structure, labels, specificity and explicit hierarchies of knowledge for signaling relevance to the reader. The National Health Information Infrastructure Initiative (http://aspe.hhs.gov/sp/nhii/) identifies academics and researchers as natural stakeholders, like clinicians and caregivers, in enabling better healthcare through better information sharing (National Committee on Vital and Health Statistics, 2003). Information science has much to contribute to the health information technology arena and to electronic health records in particular: their development, their maintenance, and most importantly their improvement to serve the needs of diverse users.

The problem

"The medical record is a material form of public memory," Berg (1996) writes, "a structured *distributing and collecting device*, where all tasks concerning a patient's trajectory must begin and end…" [Italics original; p. 510]. Structured distributing and collecting devices are the natural interest of information science.

This paper presents the medical record as a body for information retrieval research.

Medicine in *JASIST*

A search was performed for citations to *JASIS/T* in the Library & Information Science & Technology Abstracts (LISTA; Ebsco) from 1970 through the January issue of 2006. To identify articles which focused on medicine or health in general, citations were selected that were indexed with the subject descriptors “Medicine” or “Health” or included the keywords *medical* or *health* in their titles or abstracts.

This yielded a total of 214 citations. Of these 214, 45 were eliminated as false drops; for example, articles written by authors employed by schools of medicine but which did not have medical subjects, and one “article” listing books received for review. A further 39 were found to be duplicates imported from the Information Science & Technology Abstracts (ISTA) database provided by the same vendor. This left a remainder of 130 citations for analysis.

The author and a second coder with an MLS examined these 130 citations to determine the one descriptor of the major subject of each article. The number of descriptors assigned by LISTA indexers ranged from as few as 1 to as many as 8. Interindexer agreement by the two coders was 99% and consensus was reached easily.

These 130 LISTA descriptors were then mapped to the Medical Subject Headings (MeSH; National Library of Medicine) terms determined through the use of the MeSH Browser (http://www.nlm.nih.gov/mesh/MBrowser.html) to be the closest equivalent. The MeSH Browser utility permits searching not only of MeSH terms, but also the scope notes and annotations associated with each term. Because MeSH is the predominant controlled vocabulary for indexing biomedical bibliographic concepts, it permits the most specific description of medical subjects in one source of literature.

MeSH is also a multiaxial vocabulary; that is, a MeSH term can exist simultaneously in multiple hierarchies. For example, “Down Syndrome” is located not only in the *Nervous System Disease* tree, but also in *Congenital, Hereditary and Neonatal Diseases and Abnormalities*. For the purposes of this study, to facilitate classification of this literature for generalist readers in information science, terms were chosen from the Information Science tree (L01) of MeSH whenever possible. The L01 tree appears in its entirety in Table 1.[1]

[1] In a few cases, the only possible fit available was a bad one; *Knowledge Representation*, for example, in LISTA’s thesaurus, is acknowledged in MeSH only when it is enabled by a computer (*Knowledge Representation (Computer)*; representation of knowledge manually is not considered in the scope of this term. The two coders chose the term *Semantics* instead.

Table 1. Information Science as represented in MeSH
(Medical Subject Headings; National Library of Medicine).[2]

Information Science [L01]

- Book Collecting [L01.040]
- Chronology [L01.080]
- Classification [L01.100] +
- Communication [L01.143] +
- Communications Media [L01.178] +
- Computer Security [L01.209]
- Computing Methodologies [L01.224] +
- Copying Processes [L01.240] +
- Data Collection [L01.280] +
- Data Display [L01.296] +
- Informatics [L01.313] +
- Information Canters [L01.346] +
- Information Management [L01.399]
- Information Services [L01.453] +
- Information Storage and Retrieval [L01.470] +
- Information Theory [L01.488]
- Library Science [L01.583] +
- Medical Informatics [L01.700] +
- Pattern Recognition, Automated [L01.725] +
- Publishing [L01.737] +
- Systems Analysis [L01.906] +

Note: + symbol means that this portion of the tree is expandable.

Table 2 shows the distribution of medical subject coverage within JASIST between 1970 and January, 2006. The descriptors for six citations had no MeSH equivalents. These appear in italics. Levels in the MeSH hierarchy are indicated by a box symbol [□].

[2] *Source*: National Library of Medicine. (2006). *MeSH Browser*. Available at: http://www.nlm.nih.gov/cgi/mesh/2006/MB_cgi#TreeL01. (Date accessed: February 11, 2006).

Table 2. Medical subject coverage in *JASIST*
according to specific MeSH categorization, 1970-2006.

MeSH term[3]	**Number of citations**
Administrative agencies	1
Behavior and behavior mechanisms*	
□□□ Peer review	1
Buildings	1
Communication	1
□□ Linguistics	
□□□ Semantics	1
□□□ Vocabulary	1
□□□ Writing	1
□□□□ Authorship	1
□□Language arts	1
□Information dissemination	1
Communications media	
□ Publications	2
Computer systems	
□□ Internet	4
Computing methodologies	
□ Algorithms	1
□ Artificial intelligence	1
□ Automatic data processing	1
□□ Natural Language Processing	2
□ Computer systems	1
□ Hypermedia	1
Copying process	1
Data Collection	
□ Medical records	1
□ Surveys	1
Diagnosis*	
□□ Diagnostic imaging*	3
Government information	1
Grammar, comparative and general	1

[3] All descriptors are taken from the L01 tree of MeSH unless identified by an asterisk [*].

Table 2. Medical subject coverage in *JASIST* continued

Health care economics and organizations*	
□Congresses	1
Health care facilities, manpower, and services*	
□ Delivery of health care	1
□ Health services accessibility	1
□ □ Mental health services	1
Health care quality, access, and evaluation*	
Health occupations*	
□ Medicine	2
Information Centers	
□ Libraries	4
□□ Medical libraries	1
Information science	6
Information services	12
□ Abstracting and indexing	2
□ Cataloging	1
□ Documentation	1
Information storage and retrieval	18
□ Databases, bibliographic	4
□ Databases	5
□□ MEDLINE	2
Literature	1
Publications	
□ Book reviews[4]	6
□ □ Bibliography	6
□ □ Dissertations, academic	1
□ □ Periodicals	9
Medical informatics	5
Persons*	
□ Patients	1
Psychological phenomena and processes*	
□□Task performance and analysis	1
Publication formats*	
□ Abstracts	1
□ Letter	1
Publishing	4
Training	1
Total	130

[4] Citations to "Books Received but Not Reviewed" were not included in this total (e.g., Wallace, D.P. (1988), v. 39(1), pp. 26-27).

Categorizing the descriptors according to the most general level of the MeSH hierarchy results in the gross subject distribution shown in Table 3, below. The descriptors are presented in descending order by citation.

Table 3. Medical subject coverage in *JASIST* according to general MeSH categorization, 1970-2006.

MeSH term[5]	**Number of citations**
Information storage and retrieval	29
Communications media	25
Information services	16
Communication	7
Computing methodologies	7
Information science	6
Information Centers	5
Medical informatics	5
Computer systems	4
Publishing	4
Diagnosis*	3
Data Collection	2
Health care quality, access, and evaluation*	2
Health occupations*	2
Publication formats*	2
Administrative agencies	1
Behavior and behavior mechanisms*	1
Buildings	1
Copying process	1
Government information	1
Grammar, comparative and general	1
Health care economics and organizations*	1
Health care facilities, manpower, and services*	1
Persons*	1
Psychological phenomena and processes*	1
Training	1
Total	130

It is clear from the numbers in Table 3 that information science's predominant interest in medicine and health, at least as evidenced in this flagship journal of the field, has been

(1) information retrieval,
(2) communication of the information once retrieved, and
(3) the services that support retrieval and communication.

[5] All descriptors are taken from the L01 tree of MeSH unless identified by an asterisk [*].

This is not a surprising finding since these are three domains that would be considered predominant interests of many generalist information scientists as well. When we speak of medical information in information science, we usually are speaking of information sources and tools developed by the National Library of Medicine or its sister institute the National Center for Biotechnology Information—the codevelopers and maintainers of PubMed. For example, there are numerous articles about MEDLINE and its exploitation; Swanson's famous fish oil article (1987) is an early example, Leroy & Chen (2005) a more recent one. The Unified Medical Language System has also been featured in the pages of this journal (for example, Humphrey, Rogers, Kilicoglu, Demner-Fushman, & Rindflesch, 2006).

But of the 130 articles published about medicine in almost 36 years of *JASIST*, 70 (54%) deal with information retrieval, communication and the work processes behind them, while only 2 of these (1.5%) have focused on the medical record. The first of these was authored by William Hersh, MD, a prominent medical informatics researcher and director of the NLM Medical Informatics training program at Oregon Health Sciences University. His article appeared in 1995, and served as a forecast of technology trends to come with implications for the information science profession. The second paper appeared 6 years later and is the only *JASIST*-published study focusing on the medical record as a document set for experimentation. "An Experimental Study in Automatically Categorizing Medical Documents", presented a study of automatic classification using the International Code of Diseases (Ribeiro-Neto, Laender, & de Lima, 2001). The document base in this study consisted of over 20,000 clinical documents from a contemporary healthcare system. It is this author's contention, which will be explored in the remainder of this paper, that the medical record deserves consideration *as a document base* for information retrieval research.

The medical record

In 1965, Berkeley lamented the "chaos of medical information-gathering" and concluded that "this information is more or less useless in terms of being retrievable by computer methods":

> The usual case-history form often represents a device for recording and reinforcing the interests and prejudices of the individual physician or clinical investigator reflecting current fashions in diagnosis. It is frequently no more than an essay by the physician of doubtful literary or scientific merit (1965, p. 4).

Although Berkeley's motive was criticism of case histories and case history takers, he also calls our attention to the fundamental function of the medical record, whether paper or electronic. Throughout its development, it has always documented both the knowledge domains of clinical practice, and the work processes and practices that support and maintain the operation of these domains.

The EHR and its Contents

The Institute of Medicine, in its report *Key Capabilities of an EHR System*, stresses that the modern motivation for an Electronic Health Record, or EHR—one of many acronyms in use for an electronic health record--is not a desire for "a paperless record per se, but to make important patient information and data readily available and useable." (Committee on Data Standards for Patient Safety, 2003). For this reason, the EHR has been defined as a "complete online record that is accessible to all that need it when it is needed" (Ondo, Wagner, & Gale, 2002, p. 2).

The EHR is fundamentally a "container for a set of transactions". These transactions are both persistent ones, such as historical data pertaining to one patient, with long-term value; and records of individual events, such as EKG tracings of that same patient on one morning in a single clinic, data that has short-term value (Bird, Goodchild, & Beale, 2000). The need for longitudinal access to persistent information is one characteristic that distinguishes healthcare IT from other industries, which typically experience "heavy retrieval requirements initially and then a drop-off in the need to access records". In healthcare, conversely, "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." (Cisco, 1996).

The paper medical record has historically supported—and thus the EHR must continue to support—numerous work processes and subprocesses, with multiple authors and custodians; potential audiences; intended data lifespans; and trajectories documenting care in different locations and for different purposes. As Khare and Rifkin (1998) remind us, "Usage determines community, which in turn refines the common ontology" (p. 393). So researchers interested in exploring semistructured medical documents, both for processing and for EHR systems development purposes, have followed one of two paths: automated extraction of text from existing documents to build a new ontology, or manual analysis of existing text to build a new ontology. The point of the ontology is to use it to structure the next phase of the resource's development.

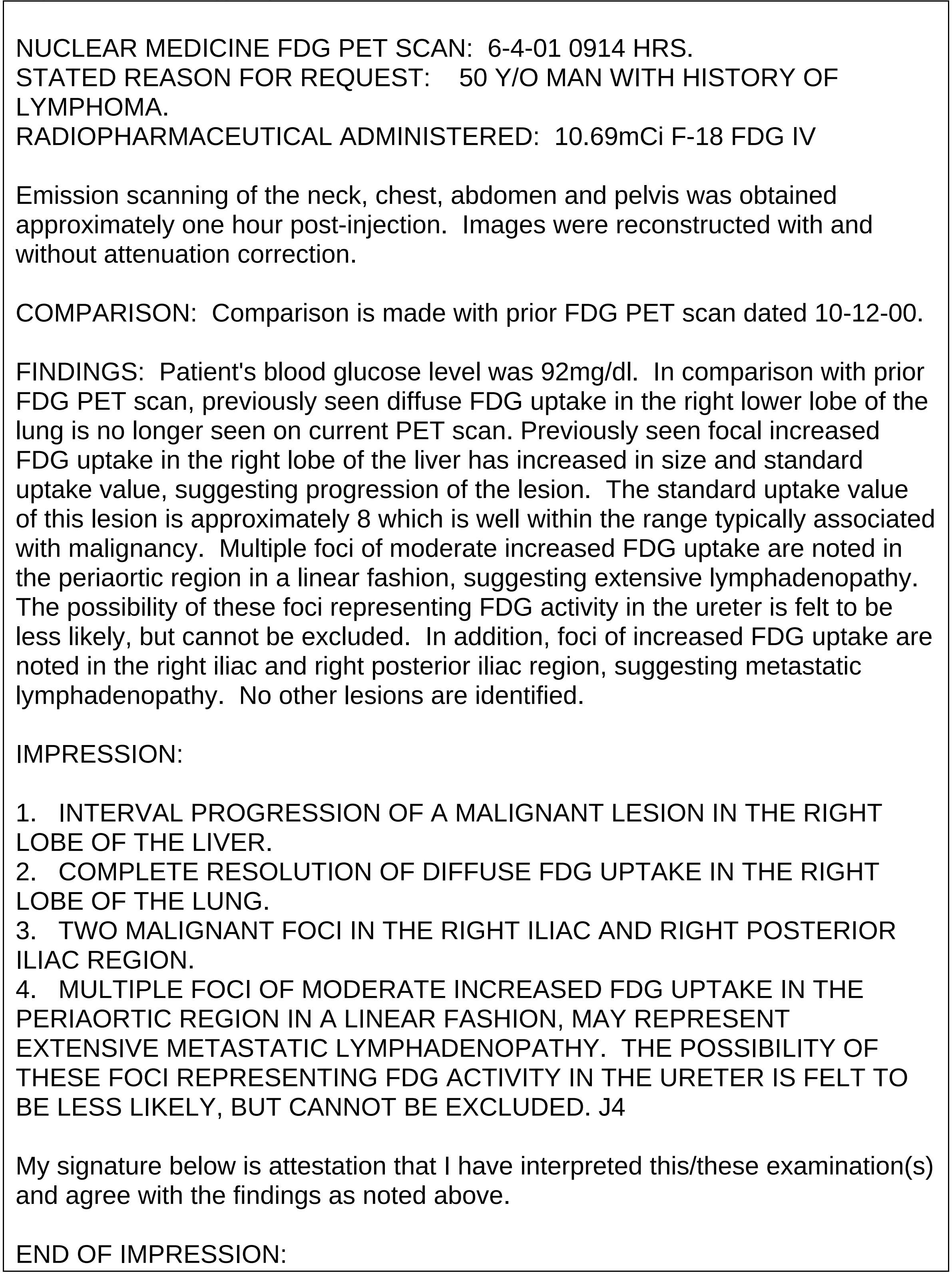
The content of EHRs reflects this multiplicity of needs and audiences. It is a mix of highly structured numeric data and excessively unstructured and idiosyncratic narrative text; increasingly, images are included as well. In fact, any information can be part of the medical record that is relevant for clinical decision making. 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 the structure in medical records, the current paper-based record has these typical contents which are present, in various degrees, in EHRs as well.

- 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 generated outside the health care organization but maintained as part of the individual patient’s history.

Much of this information can be and is presented to the user in the context of a text-based document. Clinical documents may also refer to each other. These documents are frequently “nested” inside each other, also; for example, EKG narrative reports can appear within a cardiology report; a letter from a physician may include results of a genetic test (Smith, 2005).

An example of a typical clinical document found in an electronic medical record system appears in Figure 1. Dates and ages have been pseudonymized.

Figure 1. Radiology report.



NUCLEAR MEDICINE FDG PET SCAN: 6-4-01 0914 HRS.
STATED REASON FOR REQUEST: 50 Y/O MAN WITH HISTORY OF LYMPHOMA.
RADIOPHARMACEUTICAL ADMINISTERED: 10.69mCi F-18 FDG IV

Emission scanning of the neck, chest, abdomen and pelvis was obtained approximately one hour post-injection. Images were reconstructed with and without attenuation correction.

COMPARISON: Comparison is made with prior FDG PET scan dated 10-12-00.

FINDINGS: Patient's blood glucose level was 92mg/dl. In comparison with prior FDG PET scan, previously seen diffuse FDG uptake in the right lower lobe of the lung is no longer seen on current PET scan. Previously seen focal increased FDG uptake in the right lobe of the liver has increased in size and standard uptake value, suggesting progression of the lesion. The standard uptake value of this lesion is approximately 8 which is well within the range typically associated with malignancy. Multiple foci of moderate increased FDG uptake are noted in the periaortic region in a linear fashion, suggesting extensive lymphadenopathy. The possibility of these foci representing FDG activity in the ureter is felt to be less likely, but cannot be excluded. In addition, foci of increased FDG uptake are noted in the right iliac and right posterior iliac region, suggesting metastatic lymphadenopathy. No other lesions are identified.

IMPRESSION:

1. INTERVAL PROGRESSION OF A MALIGNANT LESION IN THE RIGHT LOBE OF THE LIVER.
2. COMPLETE RESOLUTION OF DIFFUSE FDG UPTAKE IN THE RIGHT LOBE OF THE LUNG.
3. TWO MALIGNANT FOCI IN THE RIGHT ILIAC AND RIGHT POSTERIOR ILIAC REGION.
4. MULTIPLE FOCI OF MODERATE INCREASED FDG UPTAKE IN THE PERIAORTIC REGION IN A LINEAR FASHION, MAY REPRESENT EXTENSIVE METASTATIC LYMPHADENOPATHY. THE POSSIBILITY OF THESE FOCI REPRESENTING FDG ACTIVITY IN THE URETER IS FELT TO BE LESS LIKELY, BUT CANNOT BE EXCLUDED. J4

My signature below is attestation that I have interpreted this/these examination(s) and agree with the findings as noted above.

END OF IMPRESSION:

Medical Data and Work Practices

"Insufficient information" has been implicated as one of many failure modes resulting in medical errors and adverse events: insufficient meaning a lack of information regarding the drugs the patient has been prescribed; previous dose-response relationships; pharmaceutical information; laboratory data; and known allergies (Kohn, Carrigan & Donaldson, 2000). The electronic medical record, as the principal information resource for clinical care, has the potential to solve these problems through making clinical information more represent-able and thus retrieve-able.

However, considering the medical record as an information *source* requires knowledge of the work practices that the record supports. "The very *possibility* of understanding the record's entries is based on a shared, practical, and entitled understanding of common tasks, experiences, and expectations" (Atkinson & Heath, 1981, pp. 200-201). As one social historian explains:

> It is a mistake to separate the knowledge claims of medicine from its practices, institutions, and so on. All are socially fashioned, and so it may ultimately be more helpful to think of mentalities, modes of thought, and medical culture than in terms of "knowledge", which implies the exclusion of what is inadmissible. [Jordanova, 1995, p. 362].

Sociologists Garfinkel and Bittner remarked on the intertwining of practice and documentation in the 1960s when, as investigators, they attempted unsuccessfully to intervene in a medical clinic's recordkeeping practices: "Attempts to pluck even single strands can set the whole instrument resonating", they note diplomatically (1967, p. 192).

Medical data have three characteristics relating directly to work context.

Specificity and Purpose

> First, data are always produced with a given purpose, and their hardness and specificity is directly tailored to that purpose…the meaning, hardness and significance of a piece of information cannot be detached from the specific purpose that structured the gathering of that information (Berg & Goorman, pp. 53-54).

This is a sociologist's perspective on the difference between free text and controlled vocabularies. In healthcare, the "rigidity" of coding and classification exists so that information can be generalized across a number of clinical situations. Consider the International Classification of Diseases (ICD) concept "Brain Neoplasms, Miscellaneous, Not Otherwise Specified". This is a "soft" knowledge representation because it is imprecise. However, this very imprecision makes the concept portable across different clinical situations—for example, between African and European medicine—and even into domains outside of immediate clinical care, such as epidemiological reporting.

Contrast this with the much "harder" representation, "Astrocytoma", a label attached to a specific kind of brain tumor. This is a representation achievable only by the clinician who exits the classification system of ICD-9 to express the diagnosis the only way she can--in free text. This label's *purpose* is to accurately, and with the greatest precision possible, represent the clinician's diagnosis.

A clinical information system that has not allowed for the existence of an astrocytoma renders this diagnosis invisible—and unfindable, except via keyword searching. Therefore this diagnosis is unshareable, both within the dataset of patients represented in this information system, and across the system, to be analyzed with other types of data, such as surgical procedure, name of surgeon, or age of the patient. Just as the task of the statistician or third-party payer has a different purpose from the task of the clinician, so the granularity and quality of the diagnosis' representation will differ accordingly.

Mutual Elaboration

Second, medical data do not exist in isolation, but "mutually elaborate each other", as "bits and pieces of an emerging story" (Berg & Goorman, 1999, p. 54), subject to the effects of time. The context of data elements as being located *near* other data elements is thus also important. In fact, the "story" does not even have to be told in textual narrative; Description of the course of an illness, whether told in words or in laboratory test values, requires integration of individual data with the larger picture. For an illustration, refer to Table 4, below (Bergeron, 1998). This table shows the importance of context in interpretation of otherwise ambiguous notations in medical records; understanding of the vocabulary and world view of the specialty is necessary to make sense of acronyms and abbreviations. For example, as Bergeron comments, "often written abbreviations are unambiguous in the context of a specialty, such as 'rih' for 'right inguinal hernia', and 'GA' for 'General Anesthesia' in surgery and anesthesia." (p. 575).

Table 4. A Comparison of Traditional Handwritten Medical Record Entries (Left) and Their Equivalent Oral Translations (Right)

(Bergeron, 1998, p. 575).

Hand written tradition	Translation
Pt returns for F/U of HBP	Patient returns for follow-up of high blood pressure
On exam—class 4 airway, opens mouth ~5 cm, 3 fb thyromental dist, good neck extension	On exam, class 4 airway, opens mouth approximately 5 centimeters, 3 finger breadth thyromental distance, good neck extension
Uterus: NSSC, no mass	Uterus of normal size, shape, and consistency, no masses
…debridement stsg R ear…	…debridement single thickness skin graft of the right ear …
NPO p MN	NPO past midnight
…for rih repair…	…for right inguinal hernia repair…
…9 hours of GA…	…9 hours of general anesthesia…
PMH: NSVD times 2	Past Medical History: Normal spontaneous vaginal delivery times two
Chest: Clear to P&A	Chest: Clear to percussion and auscultation
RTC 2 weeks	Return to clinic in 2 weeks
…pulmonary hypertension with DOE for one year…	…pulmonary hypertension with dyspnea on exertion for one year…
WD obese BF NAD	Well developed, obese black female in no acute distress
…for an Ax block…	…for an Axillary block…
…the ett tape was under tension…	…the endo tracheal tube tape was under tension…
…to the patient's cvp and palpation…	…to the patient's central venous pressure (or cvp) and palpation…
Breasts: no mass or D/C	Breasts, no masses or discharge

Context of Production

Finally, a third characteristic of medical data is that human readers of medical information interpret and reinterpret "in the light of who generated it" (Berg & Goorman, 1999, p. 55), whether that generator is a human being or a machine. Medical readers *consciously* perceive the context of production, and integrate an understanding of the producer into their understanding of the data. Berg and

Bowker (1997) have even made the case that the data produced mirrors the organizational structure of the organization that produces it. These authors commented that in considering the electronic medical record, it was

> tantalizing to assert a connection between the databases drawn upon and the work organization … the hierarchical database echoes the hierarchical organization structure most favored in the 1960s; the relational database echoes more the team model of the 1970s; and object orientation is the nec plus ultra (*sic*) of radical outsourcing [p. 534, n. 14].

Clinical information retrieval

Hersh (1996) has identified the two distinctly different goals of general versus clinical information retrieval. The purpose of the former process is to get a particular document that matches the seeker's specific information need; the goodness of fit of document to need is facilitated through the use of descriptors representing the document's subject matter and manipulated by an automated system. The document is indexed; the document is retrieved; the document is the deliverable.

In contrast, the information need in the clinical setting is typically centered around a particular patient; whether the information-seeker's ultimate intent is to use this clinical data in isolation, or in aggregation with other data describing other patients. Thus the clinical data—"The digoxin level of patient 13 upon admission?"—is in fact the deliverable, not the document. This has tremendous implications for the accuracy and granularity of representation in these documents. A "false drop," or mismatch between query and result, in nonclinical situations constitutes only information noise; in clinical retrieval, it can literally be fatal. (For an excellent review of research into clinical information retrieval, see Mendonça, Cimino, Johnson, & Seol, 2001).

Typical clinical tasks performed using electronic medical record systems include the following (Laerum et al., 2001). These tasks illustrate the range of information needs and information retrieval features required:

- Review the patient's problems
- Seek out specific information from patient records
- Follow results of a test or investigation over time
- Obtain results from new tests or investigations
- Enter daily notes
- Obtain data on investigation or treatment procedures
- Answer questions concerning general medical knowledge
- Produce data reviews for specific patient groups

- Order clinical biochemical laboratory analyses
- Obtain results from clinical biochemical laboratory analyses
- Order x ray, ultrasound, or CT investigations
- Obtain results from x ray, ultrasound, or CT investigations
- Order other supplemental investigations
- Obtain results from other supplemental investigations
- Refer patient to other departments or specialists
- Order treatment directly (medical, surgery, or other)
- Write prescriptions
- Write sick leave notes
- Collect patient data for various medical declarations
- Give written specific information to patients
- Give written general information to patients
- Collect patient information for discharge reports
- Check and sign typed dictations

Safran and Chute (1995) have delineated four specific ways of reusing clinical data which are four general categories into which the clinical tasks described by Laerum et al. can be sorted:

> Results reporting: Displaying information about an individual patient
>
> Case-finding: Finding data about another patient similar to the current patient
>
> Cohort description: Describing a group of patients with at least one attribute in common
>
> Predictive modeling: Elucidating patterns in data in hopes of describing trends, or relationships, between attributes

The first application, *Results reporting*, centers on the patient as an individual, and this is the most common of the uses of clinical data because the record's primary function is to represent that patient's encounters with the healthcare system.

This does not mean, however, that retrieval of information about that specific individual is going to be easy.

> The most significant information management challenge posed by claims data is the fragmentation of patient information over time and geographic space as patients move through a fragmented treatment system. Despite the industry's belated compulsion toward horizontal and vertical integration, most patients still receive their care across

> myriad settings and sites… [there is an] unimaginable volume of clinical rules necessary to develop coherent “episodes of care”, a mind-bending data-handling task that must be completed before any meaningful clinical information can be developed and used. (Kleinke, 1998, p. 29).

Chronic disease is a long-term variation on the same theme; one individual with multiple visits to multiple providers presents the same clinical information retrieval problem multiplied exponentially. Chronic disease “must incorporate far more information than just physicians’ opinions, must be accessible from many different sites of care, and must capture accurately both the illness trends and their speed of change” (Holman, 1996, pp. 1-2).

The other three tasks outlined by Safran and Chute—*case-finding*, *cohort identification*, and *predictive modeling*--require aggregation of individuals into coherent subgroups; thus, identification of, and retrieval according to, common attributes is critical to the clinical research process. Possible attributes include age; gender; ethnic origin; and patient experience, which may include all of the above: for example, a clinician might be interested in the postoperative infections found in all Black men undergoing a particular procedure in the same hospital over time. The case-finding application described by Safran and Chute is a “Get me more like this” request; physicians “may find it useful to recall past situations similar to the current one, but the process is often biased by the tendency of recalling only more recent cases” (Montani & Bellazzi, 2001, p. 499), and an automated retrieval system helps avoid this human bias. The definition of useful attributes and the ability to restrict searches to patients with those characteristics is thus another key requirement in clinical information retrieval system design.

Thus, the importance of clinical data extends even beyond the limits of any individual patient’s healthcare needs. Clinical information is necessary for retrospective studies; outcomes research; quality assurance audits and evaluation (Dambro et al., 1988; Marshall, Balas, & Reid, 1997); decision support; management of patient care (Hersh, 1996); distributed health care (Dambro et al., 1988); and, in fact, any form of scientific research that requires dealing with individuals as members of groups (Sujansky, 1998).

Loosely structured documents

Essin & Essin (1990) had proposed loosely structured documents as the ideal electronic patient record implementation, integrating the high-quality representation available through standardized coding with the flexibility and customization afforded by a paper-document structure. Loosely structured, or semistructured, documents were earlier defined by Essin (1993) as documents that have much in common—enough in common that a general statement can be made about their components.

Essin’s strategy had two levels. The first consisted of understanding the data elements themselves. The second level was a “meta-level” that contained knowledge *about* the data elements. Essin’s idea was to capitalize on both kinds of knowledge to be able to treat the document as a piece of text with fields like a database; to consider the meta-knowledge, or structure, as a different entity from its text, or content, and model the two separately (Royal College of General Practitioners, 1999). Essin formally proposed this as an SGML solution. Since SGML is intended to separate the document and its tags, or notation, from its application specifics, these authors argued that the “ability to manage loosely structured formats avoids rigid formalisms that have the sole purpose of making the data processable” (Lincoln & Essin, 1995, p. 229). This SGML initiative eventually bore fruit as the ANSI standard Clinical Document Architecture from ISO Health Level Seven.

Health Level Seven, or HL7 (www.hl7.org), is an international ANSI-accredited Standards Developing Organization within the domain of healthcare, specifically concerned with standards for clinical and administrative data. The HL7 community includes not only academics and healthcare professionals, but representatives of every major vendor in the healthcare IT industry, which ensures industry input and compliance with the standards developed by the organization. For this reason, HL7 and the Institute of Medicine were the two entities charged by U.S. Secretary of Health and Human Services Tommy Thompson in 2003 with developing standards for a U.S. electronic health record.

The intent of the CDA in health informatics is similar to that of the Encoded Archival Description (EAD) in archives; both architectures attempt to impose order on semi-structured text documents by standardizing frequently occurring segments within those documents. The EAD was developed in 1993 as an encoding standard for machine-readable, sharable text created by libraries, museums, archives and manuscript repositories. The CDA, similarly, standardizes templates for radiology reports, laboratory test results, history and physical notes, discharge summaries, operating room notes, and hundreds of other common healthcare documents. Both the CDA and EAD are communication standards that specify structure, but do not attempt to define semantics of the content being structured. (See Smith, 2002, for a complete review of SGML/XML in medicine and the evolution of the Clinical Document Architecture standard and Arnott Smith, 2002, for an information retrieval experiment testing the effect of the standard; for the current status of Health Level Seven’s CDA and other standards, see HL7, 2005).

Features

What are the important semantic features of loosely structured documents in medicine, and how do they make clinical information accessible to the clinical reader? Wolff, Flörke, and Cremers (2000) point out that the principal defining feature of structured documents is the presence of explicit semantics for their structural parts. The benefit is that the meaning of the structural components—the sections—can be exploited, as can the meaning of the text they contain. The sections thus make up the “meta-level” proposed by Essin (1993). Elements, their labels, and their granularity are three important structural dimensions of clinical documents that have their own particular implications for retrieval of these texts.

These document sections can be considered from three different perspectives. First, there is the behavior of the sections themselves, considered to be distinct elements, or components, of the documents; second, there are the strings of text, section headings or “labels” by which these sections can be accessed by the reader; and third is the number of these elements into which the document has been divided, or “partitioned”. Each of these perspectives has its own implications for the structured document as a knowledge representation.

Elements

Lambrix and Shahmahri (2000) state that “the logical structure of a document can be seen as defining a part-of hierarchy over the document and its parts” (p. 290). In the same vein, European medical record standardization efforts have defined a “record item” as

> Part of a “chain” (having a “name” and a “content”) [that] is a part of a “record item complex” that in turn can be part of record item complexes of higher rank. This chain of complexes is the “context” for the record item. (Rossi Mori & Consorti, 1999, p. 132).

Document parts, also called elements or components, need to be understood as having meaning, particularly in relation to each other. This chain of relations is the “information space” which Kluge (1996) sees as connecting the dots between individual points of clinical data: “Only data-in-relation are information” (p. 253). In the example of the radiology report (Figure 1), the bulleted list of clinical “findings” is a distinct element present in this electronic medical record system’s radiology report document type, but nowhere else.

Labels

Nygren, Johnson, and Henriksson (1992) were the first researchers to examine the process of clinicians' reading of the medical record as a specific information source. These authors identified three reading techniques: first, skipping over irrelevant sections; second, skimming sections identified as possibly relevant; and third, reading needed information carefully. Tange, Dreessen, et al. (1997) collapsed the first two filtering steps into one in an EHR setting, describing a user who searches through the record "guided by the internal structure" to select relevant sections, then reading the content (p. 158).

Readers of clinical documents navigate by reading the labels assigned to the structural elements of the documents. These are variously called "section headings", "labels", and "segment labels" in the medical informatics literature (see for example Lincoln & Essin, 1995, and Tange et al., 1998). In Figure 1, some labels are "Comparison", "Findings", and "Impression".

The labels alert the reader to content (Lincoln & Essin, 1995). In this fashion the labels themselves serve to denote the structure and define the domain of knowledge:

> The structured representation acts as an intensional definition, in the particular vision of a world embedded in a structure. (Rossi-Mori, Galeazzi, Consorti, & Bidgood, 1997, p. 650).

This makes them useful even at shallow depths of detail. Rossi Mori and Consorti (1999) comment:

> Even without using further structures, document names or titles of sections may be useful .. to build a cumulative table of content from various record systems … to limit the search within entries. (p. 132)

In fact, the presence of specific labels in a clinical document may be a signature of a particular document type. The author's doctoral dissertation experiment used 8 common clinical document types from the Medical Archival Retrieval System (MARS) at the University of Pittsburgh Medical Center. For this experiment, two hundred and one individual labels were created from section headings present in document text. Of the 201 labels, not one was common to all 8 types. Seven of the 8 document types shared only four elements: "Addendum", "Impression", "Medications", and "Plan."

Granularity

The number of elements into which a document can be partitioned has been called the document's "granularity" (Tange, Dreessen et al., 1997). "Granularity" of clinical documents is "the level of detail to which these elementary paragraphs are specified" (p. 158). These researchers outlined a continuum of granularity like that of sandpaper: Coarse; Intermediate; and Fine. The radiology report in Figure 1 is of intermediate granularity because although some elements of the document have labels (such as "Findings" and "Impression"), others do not; for example, the precise type of radiology study performed is implicit in the text of the report (a PET scan) but is not explicitly called out by a label.

Passage Retrieval

The body of existing information retrieval work most relevant to the medical record as a base for experiment is the work called "passage retrieval". This has been defined as "the task of identifying and extracting fragments from large, or short but heterogeneous full text documents" (Melucci, 1998, p. 44). It is a subset of research into "corpus-based" text processing. Here, the text collection itself is used to derive information needed for analysis and for characterization (Salton & Allan, 1993). Moffat, Sacks-Davis, Wilkinson, and Zobel complained in 1994 that relatively little work had been done on retrieving or ranking partial documents (p. 188); a good review of research through the late 1990s can be found in Melucci (1998). More recent explorations include those of Fuji, Iwayama & Kando, who used a patent collection (2004) and Cui, Sun, Li, Kan & Chua (2005), who were interested in answer passages in question-answering systems.

Passage retrieval was inspired by the problem of large documents. As Melucci comments (1998):

> Large is relative; the less the power of computational resources, or the worse the system capabilities of providing the user with usable and useful access to large documents, the more the documents are to be considered as large (p. 43).

O'Connor, one of the earliest investigators of passage retrieval, used a medical resource for his dataset, but this was a medical bibliographic database: CANCERLIT, the cancer-specific augmented Medline from the National Library of Medicine. Inspired by the success of LEXIS, the legal full-text system, O'Connor asked "Why not passage retrieval for scientists?" (1980, p. 227).

Like "large", the precise definition of "passage" has varied in the literature. Yang, Maglaughlin and Newby (2001) call this one of its principal challenges: "How documents should be split into passages in order to maximize [passage retrieval's] advantageous potential is an important consideration" (p. 527). But in general, passages have been defined as being "some semantic structural feature" of the document (Kaszkiel & Zobel, 1997). Callan (1994) summarizes the passage retrieval literature as treating three different types of passages: windows, semantic, and discourse (see also Hearst & Plaunt, 1993; Lalmas & Ruthven, 1998).

Windows are sections of text defined as a certain number of words (see Callan, 1994, for an example, and Xi & Xu-Rong, 2001, for an experiment in which window size was an independent variable). *Semantic passages* are defined based on the subject or content of the text. Hearst and Plaunt (1993), for example, worked with "subtopics" denoted by the subheadings in a magazine article.

Discourse passages are based on units of textual discourse: sentences, paragraphs, and sections. It is discourse passages that put the weightiest requirements on the composer (Callan, 1994) because these passages require consistency from writers. Thus, analysis by discourse seems to work best with highly structured and edited text, such as the encyclopedia text used by Salton and colleagues in numerous experiments (Salton & Allan, 1993; Salton, Allan, & Buckley, 1993; Salton & Buckley, 1991). In fact, if text is neither highly structured nor edited, passage retrieval is more difficult than document retrieval, because:

> Any pre-defined segmentation of the text is absent, unless the text author has provided the text itself with a structure reflecting the organization of the topic which might support the retrieval of passages relevant to the topics (Melucci, 1998, p. 44).

Both subtopics and sections are understood to be visible, explicit structural features of the text that are available for semantic processing. Their presence is alerted by strings of text that are legible labels—subheadings, or section headings: "[T]ext structure can be a good approximation of topic organization" (Melucci, 1998, p. 47).

Callan (1994) and Kaszkiel & Zobel (1997) argue that because of these elements of a document's structure, all information retrieval can be viewed as a passage retrieval task—or, at least, a task of retrieving documents that have an internal structure: "Each element is a source of evidence that can be used in retrieval" (p. 302). Blair (2002) similarly states that "the determinacy of document representation" is one of the three factors most influencing content retrieval (the other two being the size of the collection and the type of the search) (p. 303).

However, Moffat, Sacks-Davis, Wilkinson, and Zobel (1994) consider it important to consider documents not only structures, but *hierarchical* structures in particular (p. 181). Panko et al. (1999) refer to these hierarchical displays as "outlines" and "suboutlines", noting that the suboutlines deal with increasing levels of specialization, while outlines, as the most general level, offer "greatest stability".

Benefits

Passage retrieval attempts to address the significant problem of full-text document retrieval which rests in the sheer size of the documents. This characteristic of full text

> may have a confounding effect: It may be large and difficult to manage, and relevant information may be widely scattered, and therefore hard for the user to extract (Tombros & Sanderson, 1998, p. 2).

Although Wolff et al. (2000) point out that passage retrieval allows the user to make very precise *queries* because of the "enhanced expressive power" it affords, passage retrieval research thus far has virtually entirely been concentrated not on the questions, but the document "answers". In a full-text system, documents that are returned as "answers" to queries pose problems; they are large and unwieldy objects and the "answer" located within the result returned may be difficult for users to extract, so that it is not really an "answer" at all (Moffat et al., 1994). Cormack, Clarke, Palmer, and To (2000) have examined the use of passages to refine queries; their Multi-Text System supports "the retrieval of passages defined at query time rather than at build time" (p. 152). Yang et al. (2001) had good results through allowing the user to select the passages; according to these researchers, the interactivity of their system renders the "arbitrary determination of passages", discussed above, a nonissue (p. 527).

Proponents argue that the primary advantage of passage retrieval rests in its ability to enhance the relevance of results returned. O'Connor noticed this in his early CANCERLIT experiments: "Passage retrieval saves users valuable time by immediately presenting them with relevant material within a document" (1980, p. 228). But opinions are mixed. Moffat et al. (1994) state that although very long and diffuse documents may give the appearance of relevance, close examination will reveal the reverse. Furthermore, when relevance ranking is done using whole documents as the input, one passage's high relevance may be completely obscured by the low relevance of the whole document overall (Kaszkiel & Zobel, 1997).

It may be the revelation of document structure through hierarchies and outlines of the content that can best help the user determine relevance of a passage (Salton & Buckley, 1991). As Lalmas and Ruthven (1998) put it, often only *part* of a document *is* relevant. Callan (1994) states that the problem with full-text retrieval using "long documents, documents with complex structure, and even short documents summarizing many subjects" (p. 302) is that this presents a challenge when the user can't tell where in the text an answer lies. As Kaszkiel and Zobel (1997) said: "Since passages are relatively short, they embody locality." (p. 178).

Passage retrieval helps because it concentrates the reader's attention on those parts of the text that have a "high density" of relevant information, while also giving the reader an "intuitive overview" of the way in which those relevant subsections are distributed throughout the corpus (Tombros & Sanderson, 1998, p. 2; see also Salton et al., 1993; Salton & Allan, 1993). This phenomenon was also noted early by O'Connor (1980). In one experiment, he obtained questions about cancer "from those used at a CANCERLIT terminal" (perhaps log files, but not directly stated; p. 228) and directed medical librarians to construct an answer-base, using any means *except* CANCERLIT to answer the question. In two cases, answers could not be found; a computer then selected passages from the full-text articles that contained answers to the rest. O'Connor found that answers tended to be located in particular places, and that this knowledge could be exploited for ranking purposes:

> Another interesting result is the following: About 20% of the falsely retrieved sentences .. were from sections of papers which were headed "Methods" or "Materials and Methods". Another 20% were from sections headed "Discussion". … By ranking output passages from such sections lower than those from the rest of the paper, nearly the same recall … could be achieved with a 40% reduction in false retrieval. (p. 236)

Losee (1996) found similar results in another medical domain: the CF database within TREC, which consists of Medline citations to articles to which the MeSH term "Cystic Fibrosis" was attached. (For a description, see Shaw, Wood, Wood & Tibbo, 1991). Working with window passages, Losee hypothesized that text "windows" and phrases would differ according to the knowledge discipline from which they were drawn. His experimental results showed that the "statistically significant window" could be used as a distinguishing characteristic of text corpora, although the reasons Losee cites may surprise anybody who has worked with medical terminologies:

> Authors of articles in the medical literature use the terms in the abstract more frequently in the body of the text and with more regularity than is found in other disciplines. This may be due both to a consistent and unambiguous vocabulary for medical discussions…(p. 755).

Salton et al. (1993), who "scored" sentences in ranking of passages, note that "the location of sentences in the text under consideration" affects a particular passage's ranking (p. 50). Salton and Allan (1993) point out too that an understanding of local context allows the user and the system to avoid false retrievals “caused by linguistic ambiguities” which context obviously disambiguates (p. 132). Context is also important when considering the relationship of document components *to each other*, since these components will relate “both temporally and hierarchically” (Lalmas & Ruthven, 1998, p. 530). These improvements in delivery of *relevant* documents to the user may be due to the ability of passage retrieval systems to exploit user understanding of *context* in relevance assessment.

Clinical documents as candidates for passage retrieval

The author has been unable to identify any acknowledgment of the passage retrieval research in the medical informatics literature or vice versa. One principal medical research group ascribes to “structured retrieval” all the same advantages of passage retrieval. This group is located at Maastricht University in the Netherlands, and is headed by Huibert Tange (see Tange, Dreessen et al., 1997; Tange, Hasman, Robbe, & Schouten, 1997; Tange, 1996; Tange et al., 1998). Tange, Hasman et al. (1997) propose that the “search structure”, or information retrieval process, in the domain of clinical information has two main aspects: the “granularity” of the paragraphs, as previously defined, and the relationship of those paragraphs to each other. The number of paragraphs being searched—an important aspect of granularity--is proposed to be inversely related to the ease of searching them, which relates obviously to the work of Salton et al. ascribing passage understanding to relevance. The same proposition was put forward by Lincoln and Essin in 1995, again with only implicit acknowledgment that relevance had anything to do with the desired result: “An ability to specify text searches as narrowly as necessary using additional tags [SGML] avoids secondary parsing or sorting *to eliminate unwanted material*” (italics by author; p. 229).

In one experiment, Tange et al. (1998) found that high granularity of clinical documents (meaning documents with large numbers of sections) was associated with increased speed of information retrieval for progress notes only; certainly, a finding of high value in a high-need clinical situation. However, this finding did not hold for other types of documents, specifically Medical History or Physical Examination documents, where excessive partitioning caused more problems than it solved. The author’s doctoral dissertation, which tested information retrieval in 1000 structured versus unstructured clinical documents, found that structuring had no effect on precision (Smith, 2002). These findings agreed with those of Moffat et al. (1994). The effect of document type on the passage retrieval process clearly requires further research.

In addition, results from the relatively small body of passage retrieval literature to retrieval of structured documents in medicine—an even smaller body--need to be considered in light of the nature of the documents themselves. Salton and his colleagues (Salton & Allan, 1993; Salton & Buckley, 1991; Salton et al., 1993) in their ranking algorithms concentrated in *similarity* of passages within documents, focusing on finding “subparts of a large document” (an electronic encyclopedia) that “co-refer or are very similar in content” (Hearst & Plaunt, 1993, p. 61). This was done because similarity of passages was used as a signal of the relevance of one passage to another:

> An attempt was made to recognize text portions within which text meanings are homogeneous (that is, sufficiently similar to conclude that the texts are closely related (Salton & Buckley, 1991, p. 22).

Clinical documents are typically extremely short. The author’s doctoral dissertation used 1000 reports from a long-running electronic medical record system at the University of Pittsburgh. These documents consumed no more than 2 pages of printed text on average, were often only a few paragraphs long, and featured unique and nonredundant text organized akin to Hearst and Plaunt’s “subtopics” (1993). In addition, unlike the case of the magazine articles mined by Hearst and Plaunt, the section heading labels are not content summarizations of the paragraphs they signal. The section headings used in clinical documents are more like fields of a database than the discourse-structured text of passage retrieval experiment.

Finally, since each clinical document represents only one patient and one clinical event (e.g., a PET scan for the 50-year-old male represented in Figure 1), similarity of passages will probably seldom occur *within* documents. Instead, it is likely to occur *across* aggregations of documents that are of the same or similar type, e.g., all PET scans performed on 50-year-olds in all clinics in the healthcare system. Since, as previously described, three of the four principal types of clinical queries rely on identification of subgroups by common attributes, passage similarity is still an important factor to consider.

Conclusion

This paper has outlined a document-centered approach to the EHR as an information retrieval problem. It is clear that passage retrieval researchers working in the field of information science have seen similar values in document passages as have researchers in medical informatics. Without either literature acknowledging the other, workers in both camps have identified the same potential in document structure, labels, specificity and explicit hierarchies of knowledge for signaling relevance to the reader.

The National Health Information Infrastructure Initiative (http://aspe.hhs.gov/sp/nhii/) identifies academics and researchers as natural stakeholders, like clinicians and caregivers, in enabling better healthcare through better information sharing (National Committee on Vital and Health Statistics, 2003). Information science has much to contribute to the health information technology arena and to electronic health records in particular: their development, their maintenance, and most importantly their improvement to serve the needs of diverse users.

References

Arnott Smith, C. (2003). Effect of XML markup on retrieval of clinical documents. *Proceedings/AMIA* :614-8.

Berg, M., & Bowker, G. (1997). The multiple bodies of the medical record: Toward a sociology of an artifact. *Sociological Quarterly, 38*(3), 513-537.

Berg, M., & Goorman, E. (1999). The contextual nature of medical information. International Journal of Medical Informatics, 56, 51-60.

Bergeron, B. (1998). The effect of technology on the written tradition of medicine. *Perspectives in Biology and Medicine, 41*(4), 572-578.

Bird, L.J., Goodchild, A., & Beale, T. (2000). Integrating health care information using XML-based metadata. *HIC 2000*. Retrieved February 12, 2006, from http://titanium.dstc.edu.au/papers/HIC2000.pdf.

Blair, D. C. (2002). The challenge of commercial document retrieval, Part II: A strategy for document searching based on identifiable document partitions. *Information Processing & Management, 38*(2), 293-304.

Callan, J. P. (1994). Passage-level evidence in document retrieval. *Proceedings of ACM SIGIR Conference* (pp. 302-310). New York: ACM.

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

Committee on Data Standards for Patient Safety, Institute of Medicine. (2003). *Key Capabilities of an Electronic Health Record System: Letter Report*. Washington, DC: National Academies Press. Retrieved February 12, 2006, from http://fermat.nap.edu/catalog/10781.html].

Cormack, G., Clarke, C., Palmer, C., & To, S.S.L. (2000). Passage-based query refinement. *Information Processing & Management, 36*(1), 133-153.

Cui, H., Sun, R., Li, K., Kan, M-Y., & Chua, T-S. (2005). Question answering passage retrieval using dependency relations. *SIGIR Forum, 39*(2), 400-407.

Dambro, M.R., Weiss, B.D., McClure, C.L., & Vuturo, A.F. (1988). An unsuccessful experience with computerized medical records. *Journal of Medical Education, 63*(8), 617-623.

Essin, D. J. (1993). Intelligent processing of loosely structured documents as a strategy for organizing electronic health care records. *Methods of Information in Medicine, 32,* 265-268.

Essin, D. J., & Essin, C. D. (1990). Computerizing medical records: Software criteria for systems to document patient encounters. *Critical Care Medicine, 18*(1), 100-102.

Fuji, A., Iwayam, M., & Kando, N. (2004). The patent retrieval task in the Fourth NTCIR Workshop. *SIGIR Forum*, 560-561.

Garfinkel, H., & Bittner, E. (1967). "Good" organizational reasons for "bad" clinic records. In H. Garfinkel (Ed.), *Studies in ethnomethodology* (pp. 186-207). Englewood Cliffs, NJ: Prentice-Hall.

HL7 [Health Level Seven.] (2005, Oct. 13). HL7 position statement on CCR/CDA Harmonization. Retrieved February 12, 2006, from http://www.hl7.org/documentcenter/public/mou/HL7%20Position%20Statement%20on%20CCR%20updated%20Oct%2014.pdf.

Hearst, M. A., & Plaunt, C. (1993). Subtopic structuring for full-length document access. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 59-68). New York: ACM.

Hersh, W. R. (1996*). Information retrieval: A health care perspective*. New York: Springer-Verlag.

Hersh, W.R. (1995). The electronic medical record: Promises and problems. *Journal of the American Society for Information Science,* 46(10), 772-776.

Humphrey, S.M., Rogers, W.J.’ Kilicoglu, H., Demner-Fushman, D., & Rindflesch, T.C. (2006). Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science & Technology,* 57(1), 96-113.

Jordanova, L. (1995). The social construction of medical knowledge. *Social History of Medicine*, 8(3):361-381.

Kaszkiel, M., & Zobel, J. (1997). Passage retrieval revisited. *SIGIR '97* (pp. 178-185). New York: ACM.

Kleinke, J. (1998). Release 0.0: Clinical information technology in the real world. *Health Affairs, 17*, 23-38.

Kluge, E.-H. W. (1996). The medical record: Narration and story as a path through patient data. *Methods of Information in Medicine, 35*, 88-92.

Laerum, H., Ellingesen, G., & Faxvaag, A. (2001). Doctors' use of electronic medical records systems in hospitals: Cross sectional survey. *BMJ, 323*(7325), 1344-1348.

Lalmas, M., & Ruthven, I. (1998). Representing and retrieving structured documents using the Dempster-Shafer theory of evidence: Modelling and evaluation. *The Journal of Documentation, 54*(5), 529-565.

Lambrix, P., & Shahmehri, N. (2000). Querying documents using content, structure and properties. *Journal of Intelligent Information Systems, 15*, 287-307.

Leroy, G., & Chen, H. (2005). Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. *Journal of the American Society for Information Science & Technology,* 56(5), 457-468.

Lincoln, T. L., & Essin, D. J. 1995. A document processing architecture for electronic medical records. *Medinfo,* 227-230.

Kohn, L.T., Corrigan, J.M., & Donaldson, M.S. (Eds.) (2000). *To Err Is Human: Building a Safer Health System*.. Washington, DC: National Academies Press.

Losee, R. (1996). Text windows and phrases differing by discipline, location in document, and syntactic structure. *Information Processing & Management, 32*(6), 747.

Marshall, J., Balas, E. A., & Reid, J. C. (1997). Technique for efficient information retrieval in outpatient systems. *Journal of the American Medical Informatics Association*, 76-80.

Melucci, M. (1998). Passage retrieval: A probabilistic technique. *Information Processing & Management, 34*(1), 43-67.

Mendonça, E. A., Cimino, J. J., Johnson, S. B., & Seol, Y. H. (2001). Accessing heterogeneous sources of evidence to answer clinical questions. *Journal of Biomedical Informatics, 34*(2), 85-98.

Moffat, A., Sacks-Davis, R., Wilkinson, R., & Zobel, J. (1994). Retrieval of partial documents. In D. Harman (Ed.), *Proceedings of the Second Text Retrieval Conference (TREC-2)* (pp. 181-190). Washington, DC: Department of Commerce/NIST.

Montani, M., & Bellazzi, R. (2001). Intelligent knowledge retrieval for decision support in medical applications. *Medinfo*, 10(Pt 1), 498-502.

National Committee on Vital and Health Statistics (2001). Information for Health: A Strategy for Building the National Health Information Infrastructure. Retrieved February 12, 2006, from http://www.ncvhs.hhs.gov/nhiilayo.pdf.

Nygren, E., Johnson, S., & Henriksson, P. (1992). Reading the medical record: I. Analysis of physicians ways of reading the medical record. *Computer Methods and Programs in Biomedicine, 39,* 1-12.

O'Connor, J. (1980). Answer-passage retrieval by text searching. *Journal of the American Society for Information Science, 31*(4), 227-39.

Ondo, K., Wagner, J., & Gale, K. (2002). The electronic medical record: Hype or reality? *Journal of Healthcare Information Management*, *17*(4): p. 2.

Panko, W., Silverstein, J., & Lincoln, T. (1999). Technologies for extracting full value from the electronic patient record. *Proceedings of the 32nd Hawaii International Conference on System Sciences*. New York: IEEE.

Rees, C. (1981). Records and hospital routine. Atkinson P, & Heath C (Eds.), *Medical work: Realities and routines* (pp. 55-70) . Farnborough, England: Gower.

Ribeiro-Neto, B., Laender, A.H.F.d.L., & Luciano, R.S. (2001). An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science & Technology,* 52(5), 391-401.

Rossi Mori, A., & Consorti, F. (1999). Structures of clinical information in patient records. *Proceedings/AMIA Annual Symposium,* 132-6.

Royal College of General Practitioners Health Informatics Task Force. (1999). *Scoping the primary care view of the option for the development of the EPR (ScopeEPR).* Retrieved February 11, 2006, through Google.com cache: http://www.schin.ncl.ac.uk/rcgp/scopeEPR/report/index22.htm.

Salton, G., & Allan, J. (1993). Selective text utilization and text traversal. *Hypertext '93* (pp. 131-144). ACM.

Salton, G., & Buckley, C. (1991). Automatic text structuring and retrieval: Experiments in automatic encyclopedia searching. *Proceedings of the 14th International SIGIR.* (pp. 21-30). New York: ACM.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *SIGIR '93* (pp. 49-55). ACM.

Shaw, W. M., Jr., Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research, 13*, 347-366.

Smith, C.A. (2005). *Electronic health records: Sharing knowledge while preserving privacy*. IBM Center for Healthcare Management Research Report. Retrieved February 12, 2006, from http://www-1.ibm.com/services/us/bcs/html/chm_pubs.html.

Smith, C.A. (2002). The Clinical Document Architecture: XML semantic markup for enhanced clinical information retrieval. (Doctoral dissertation, University of Pittsburgh, Pittsburgh). [Dissertation Abstracts Online, 64/01A , 9.]

Sujansky, W. (1998). A document-centric electronic medical record system with database-centric reporting capabilities. In *Toward an Electronic Patient Record '98: Proceedings* (Vol. I) (pp. 398-403). Newton, MA: Medical Records Institute.

Swanson DR. Two medical literatures that are logically but not bibliographically connected. JASIS 1987; 38: 228-233.

Tange, H. (1996). How to approach the structuring of the medical record? Toward a model for flexible access to free text medical data. International Journal of Bio-Medical Computing, 42, 27-34.

Tange, H. J., Dreessen, V. A. B., Hasman, A., & Donkers, H. H. L. M. (1997). An experimental electronic medical-record system with multiple views on medical narratives. Computer Methods and Programs in Biomedicine, 54, 157-172.

Tange, H. J., Hasman, A., Robbé, P. F. d. V., & Schouten, H. C. (1997). Medical narratives in electronic medical records. International Journal of Medical Informatics, 46, 7-29.

Tange, H. J., Schouten, H. C., Kester, A. D. M., & Hasman, A. (1998). The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association, 5*, 571-582.

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *SIGIR '98* (pp. 2-10). New York : ACM.

Wolff, J. E., Flörke, H., & Cremers, A. B. (2000). Searching and browsing collections of structural information. In Proceedings of the IEEE Advances in Digital Libraries, pp. 141-150. New York: IEEE.

Xi, W., & Xu-Rong, R. (2001). Incorporating window-based passage-level evidence in document retrieval. *Journal of Information Science, 27*(2), 73-88.

Yang, K., Maglaughlin, K., & Newby, G. (2001). Passage feedback with IRIS. *Information Processing & Management, 37*(3), 521-.