Shredding Paper:
Moving Medical Records to the Digital Age

Graham Kaemmer
Advisor: Ruzica Piskac

December 22, 2015

Abstract

Electronic data has become the standard in health information storage, but paper remains the standard for exchange. Despite access to full records of discrete data, doctors and their staffs still use mail and fax to transmit records between institutions. The cause is a lack of appreciation for electronic data, and the effect is that data is irreparably mangled when it leaves an institution. I explain the technical challenges regarding extracting electronic data from paper, and I argue that with the paper standard, recovery of discrete data is impossible. I present a possible solution, PatientBank, which hopes to use a bottom-up network to push adoption of electronic exchange.

1 Introduction

Despite large steps in the direction of electronic medical records (EMRs), paper remains the standard for exchanging records between health care providers. This report will examine both the causes and the adverse effects of paper as a standard. I will argue that the system is in a state of gridlock—current incentives for both hospitals and EMR companies perpetuate the problem. Until there exists a better way of transmitting electronic records between providers, both research and the patient experience will be stunted.
In section 2, I will detail the current process in use at a large research hospital group, the Yale-New Haven Health System (YNHHS). I will describe their adoption of an EMR and describe how and why paper records are still the default for information exchange.

Section 3 will examine the industry’s progress away from paper, focusing on emerging standards for health information storage.

Section 4 will discuss possible workarounds, focusing on my employer, PatientBank, a company that gathers and stores medical records in the cloud. PatientBank’s goal is electronic data, and we face the same challenge as the rest of the system: paper records exchange. I will examine our efforts to extract data using OCR on paper records, but I will conclude that OCR is infeasible for extracting rigid, formatted electronic data.

I will conclude in Section 5 that a wholly different system of exchange must be adopted in order to digitize all records—in transit and at rest. I will show how PatientBank plans to do so, and why this radical change will overcome the challenges of health care IT.

1.1 Paper vs. Electronic Records

In this report, I will focus on the differences between paper and electronic medical records. Although both can be very diverse, I should draw a clear distinction between the two.

Electronic data, which I also refer to as “discrete” or “digital” data, is any medical data stored in a machine-readable format. Its “values” must map to computer primitives, such as numbers or lists or dictionaries. Any medical concept, such as the “Common Cold”, should reference a digital dictionary of medical terms, such as SNOMED\(^1\) or LOINC\(^2\). Moreover, electronic data is dependent on structure. A computer should be able to look at a set of discrete medical data and decide who it concerns, when it was gathered, and where it was gathered. There are many formats for storing medical data—each EMR company designs its own, and standards organizations design even more. I will discuss two standard formats, and their adoption, in section 3.

---

\(^1\)IHTSDO. SNOMED-CT. URL: http://www.ihtsdo.org/snomed-ct.

\(^2\)LOINC. URL: http://loinc.org.
Paper records are the alternative to digital data. They are records that have been flattened onto a series of pages. Through the flattening process, they have become a jumble of text, figures, and charts. Often (though not always), faxing or scanning removes document metadata and converts the text into images. They are, as section 2 will explain, still found to be useful in the health system: every records company can export to paper records (by printing), and every doctor can read and understand them. In this way, paper is the ultimate standard for doctor-to-doctor communication.

It is important to point out that there are electronic ways of storing paper records, such as PDF or Microsoft’s .doc format. Despite the industry’s reliance on fax as an exchange mechanism, it is increasingly popular to use electronic faxing services that act as PDF mailboxes, and some institutions even consider PDFs to be “electronic copies” of records. In this paper, I include PDF and .doc records in the category of “paper records”, because they are similarly flattened and obfuscated.

2 The Current System: An Example

2.1 YNHHS

From 2011 to 2014, the Yale-New Haven Health System (YNHHS) spent an estimated $250 million adopting an EMR built by Epic Systems. The hospital system, made up of three large southern Connecticut hospitals, cited patient access, a standardized workflow, and “improving communication” as primary motivations for integrating the electronic system. As of 2015, the integration is now complete. Patient records created at the hospitals are stored electronically and are accessible to other YNHHS physicians and researchers. Patients can view their lab results, immunizations, and diagnoses online, but only if that data was measured or created at a YNHHS hospital.

Their records from other health systems remain inaccessible. “Outside records”, such as those from a referring provider or primary care physician, remain in PDF format on YNHHS servers (if they are present at all), hid-
den from patients and too obfuscated for researchers. I will discuss the obfuscation in more detail in section 4.

2.2 Records Exchange: Intake

The problem lies in the exchange of records between unaffiliated institutions. This section will outline the current system. Let’s assume a patient, Bob, wishes to see a doctor, Alice, at a YNHHS hospital. Alice needs Bob’s medical record before his appointment. The steps for her to gather Bob’s record are as follows:

1. Alice calls Bob to ask where he has received care.

2. If Bob has only been seen at YNHHS institutions in the past, then there is no need to exchange any records. They already exist in the YNHHS EMR. This is the ideal case, and exemplifies the motivation behind YNHHS’ adoption of Epic.

3. If Bob has been seen at one or more institutions outside of YNHHS, Alice must request information from those institutions.

4. To allow Alice to request his information, Bob must send to Alice (via fax or mail) a “release authorization” form, which he signs.

5. To each outside institution, Alice faxes a “records request” form, complete with Bob’s release authorization.

6. The outside institutions “fulfill” the requests by faxing or mailing back a printed-out paper record.

7. Once Alice has received the records via fax or mail, she scans them and uploads PDFs to the EMR, where she can view them as an “attachment” to Bob’s chart.

This process, called intake, is standard in the health care system, at least among hospitals who are unaffiliated. A hospital outside of YNHHS is totally unable to insert or retrieve electronic data from a hospital inside of it. Because of this, all records must be converted to PDFs to be exchanged.
2.3 The Gridlock

Doctor Alice, despite having to spend time gathering Bob’s record, has very little reason to change anything about the system. Before she sees Bob, she has to read over his previous medical records, regardless of their format. If they were electronically formatted, the process might be a little easier, or at least more consistent. But it wouldn’t save enough time to justify building a custom electronic integration with all of Bob’s providers. Doing so is not only expensive, but also suboptimal for Alice. Although there are standard formats for transmitting electronic data, health exchange researchers have found that it remains hard to exchange electronic patient data in “robust, meaningful” ways.\footnote{Report Finds Paper-Based Strategies Still a Critical Element to Successful Healthcare Information Exchange. HIMSS. URL: \url{http://www.himss.org/News/NewsDetail.aspx?ItemNumber=22482}.} In other words, the electronic standards in place lack some of the clinical robustness and readability of paper exports.

Because exchange is operated by the doctors and their staffs, not IT departments, exporting patients’ charts as PDFs is the most widely used standard in health care.

Unfortunately, it is not a standard that facilitates research. Let’s say a researcher, Rick, wants to analyze Bob’s medical history (along with that of thousands of other patients). Because Bob’s records can only be transmitted via paper, Rick cannot access an electronic copy of any of Bob’s data, such as diagnoses and lab results. Other than by periodically surveying Bob about his health, Rick is forced to read and parse paper records to gather any discrete data for his study.

Rick’s predicament demonstrates the true problem with the current system: he is powerless to change how Alice gathers records, and although it is only a minor inconvenience for her, it totally undermines Rick’s opportunities to use the valuable electronic data.

Obviously, the system needs a better standard: a portable system for representing discrete data with arbitrary levels of detail.
3 Standardization: HL7, CDA, and FHIR

Despite the apparent lack of progress in the health exchange space, there are potential solutions waiting to be utilized. Since the introduction of electronic records, people have sought standards for storing and exchanging medical data. The most prominent standards are set forth by Health Level Seven (HL7), a not-for-profit organization devoted to standardizing health care. These efforts can be described as top-down solutions—they provide hospitals and health companies the means to integrate with each other, but they do not require or otherwise incentivize adoption.

3.1 Clinical Document Architecture

Clinical Document Architecture (CDA) is a standard for representing a variety of medical information in XML format. CDA files are designed to be transmitted between institutions, as a bridge between other proprietary data formats, such as those developed by EMR companies like Epic and Cerner.

CDA documents can be used to exchange different types of data. For example, a “Patient Discharge Summary”, or a “Pathology Report”. Their data model is based on the HL7 Reference Information Model (RIM), which is an ontological framework for expressing medical concepts.

The US Government’s HITECH act of 2009, designed to push the health care industry towards interoperability, requires that electronic health record companies (such as Epic) allow data to be exported in a C-CDA format, which is a subset of the Clinical Document Architecture.\(^5\) Despite this requirement, adoption of the CDA format has been slow. As of 2015, the law only requires that the standard be used for 50% of transfers or referrals.\(^6\) This means that a referral can never be guaranteed a C-CDA document. Without that guarantee, doctors like Alice fall back on requesting paper records. In the future, there may be a tipping point, when enough doctors


tors and hospitals expect a C-CDA to be available. Until everyone has that expectation, the network of C-CDA record exchange grows only very slowly.

There have also been complaints that CDA, or other formats based on RIM, are unsuited for the future of health care IT. Software developers, in particular, have found that the formats are not conducive to building health IT applications, because the format mixes “data” and “narrative” in its core resources. For example, in a CDA, a diagnosis is not a “sickness” per se; rather, it is an action taken by the physician. Similarly, a vital sign is not “data”; it is the “narrative” of the doctor measuring that vital sign.7

3.2 Fast Healthcare Interoperability Resources

The flaws with CDAs have contributed to the slow adoption of the CDA standard, and since 2010, HL7 has sought to replace CDA with a standard that is more data-focused: Fast Healthcare Interoperability Resources, or FHIR.8

FHIR aims to be easily understood by both clinicians and developers. Compatible with modern web standards, such as JSON and REST, the FHIR standard is much easier to integrate with existing applications. HL7 hopes that developers will push the industry towards critical adoption of FHIR.

FHIR is built around the concept of a Resource. As of October 2015, there are 93 different types of resources, such as Patient, Encounter, or MedicationOrder. The resources are designed to be served by a RESTful API. For example, a client application can request a certain patient to see their MRN. Or it can request all MedicationOrders for that patient. The protocol is much more interactive than CDA, which is designed as more of a “report”: a portable file for one-time transmission of a patient’s information. The RESTful nature of FHIR matches that of most modern web systems, greatly easing the integration process. The concept of “creation” is also

---


built into FHIR’s core, which is promising for the future of health exchange. For example, FHIR makes it possible for a vendor application to insert an “Encounter” at Hospital A into the system of Hospital B, then add observation data to that encounter after the fact. This is in contrast to the CDA standard, where the vendor would have to package all encounter details into a HL7 message and send that over a VPN connection to Hospital B.

FHIR is a promising new standard, and its focus on developer adoption will pave the way to a more application-focused future. However, in order to fix health information exchange, there are still boundaries. Hospitals must have relationships with all other hospitals from which they extract electronic data. This has to do with extensive legal requirements, which state that two entities must sign a data agreement before they are allowed to communicate any protected data. This is a complicated topic, but the implication is clear: for $N$ hospitals to form a digital exchange network, there are on the order of $N^2$ agreements that must be signed.\footnote{HIE systems skirt this requirement, but because they are always location-based, most of these legal relationships are still required for disconnected institutions.} The solution to this $N^2$ problem is a centralized system of medical record storage.

4 PatientBank: A New Hope

PatientBank is a company for gathering and sharing medical records, with the goal of becoming a “single source of truth” for the health care system. We aim to give patients their full medical records, which they can share as they wish. Extending the example from section 3, Alice would be able to use PatientBank to request Bob’s record directly from Bob, regardless of where he had received care. This is convenient for both Alice and Bob, but also for Rick, who wishes to study a large medical dataset.

4.1 History of PatientBank

PatientBank started as a simple personal health record. It allows patients like Bob to easily “order” their records using the same types of “record requests” that Alice used in section 2. When Bob’s record requests are...
fulfilled, his documents are added to his PatientBank record, which he can download or share with other providers.

Like the exchange system, PatientBank is currently based on paper records. We want to change this, and we investigated multiple ways to grow our database of electronic records. Our platform is very effective for gathering paper records, and as I wrote in section 2, requesting electronic data is almost impossible, as it is rare for institutions to support electronic exchange.

We have a large bank of paper records and no easy way to request electronic ones, so we set out to extract discrete data from the paper itself.

4.2 Automated Data Extraction

In the 2014-2015 year, two PatientBank researchers, Mikayla Thompson and Abhishek Chandra, explored methods of extracting discrete data from PDF medical records.\textsuperscript{10,11} Together, their work aimed to convert PDFs into structured formats and transform those structured formats into a proprietary PatientBank format called the Health Graph, developed by Paul Fletcher-Hill.\textsuperscript{12}

Despite months of research, little progress was made towards actually formalizing any PDF data. This fall, I set out to test why, exactly, the process is so hard.

To examine Thompson’s efforts in image analysis, I looked to Tesseract\textsuperscript{13}, a sophisticated open-source optical character recognition (OCR) program. Using realistically degraded (scanned and faxed) PDF exports of Epic data, I attempted to determine whether OCR data extraction was truly feasible. The files that I used were exported from the Epic EMR in use at Yale Health, a student health center affiliated with YNHHS. Because they came from the same electronic system, the documents were of a relatively standard

\textsuperscript{10}Mikayla Thompson. “Text and Structure Extraction from Medical Record Document Images”. In: (2015).
\textsuperscript{11}Abhishek Chandra. “Transforming XML Medical Records for Integrating Patient Data”. In: (2014).
\textsuperscript{12}Paul Fletcher-Hill. “The PatientBank Health Graph”. In: (2014).
\textsuperscript{13}tesseract-ocr/tesseract. GitHub. URL: \url{https://github.com/tesseract-ocr/tesseract}. 
format—seemingly ideal for automated data extraction.

In contrast to Thompson’s optimistic conclusion, I found that OCR was far too limited to extract truly discrete data. Figure 1 reveals the extent to which data is obfuscated during printing.

![Figure 1: Different lists of “Vital Signs” from two different Yale Health printouts](image)

Both examples in Figure 1 are “Vital Signs” measured during encounters with Yale Health physicians. The boxed regions are sections of a PDF page, and the text below them is the result of running Tesseract on them. Obviously, this is not an optimized system, as Thompson’s was, but Figure 1 demonstrates the general unreliability of OCR. Discrete records rely on being exact—interpreting a blood pressure of “128/74” as “1251741” is not only wrong, but potentially dangerous. Surely the system could be made to be more accurate, but there will always be some margin of error.

Furthermore, an automatic PDF reader will always have to make assumptions about the structure of the data. In these two examples, fields are separated not by table walls, but by vertical bar characters (which Tesseract consistently confuses with “1”s and “l”s). In other parts of the document, immunizations are in table format, and medications are compiled into a bulleted list. Section titles are inconsistently justified, and tables’ headers are split onto separate pages.

Our goal at PatientBank is to build a central record bank, from which
all hospitals can request records. To do so, we must be able to provide standard data formats in order to lower integration costs. It is by no means impossible to extract some data from paper records, but it is unrealistic to expect that we can parse PDFs into standard data formats such as FHIR.

5 The Future of Exchange

The solution to the paper record gridlock is a new system of records exchange. This is not a secret. Governments have been building state-wide health information exchanges (HIEs) for years, in the hopes that their adoption will reach the aforementioned “tipping point”, where enough institutions use a platform for it to be effectively relied upon. A 2014 study found that despite HITECH requirements, only 30% of hospitals nationwide use HIEs to exchange records with unaffiliated institutions. This is nowhere near high enough to be relied upon. Other solutions, particularly ones that work on a small scale as well as a large one, are perhaps more feasible.

5.1 Top-Down is Flawed

The government-mandated HIEs rely on top-down rules and regulations to push adoption. This type of incentive is insufficient, as it lacks backwards compatibility.

Let’s return to Doctor Alice, and let’s say that her hospital participates in an HIE. Because she (or her staff) is in charge of gathering Bob’s record, she has a choice between requesting Bob’s paper records and requesting his electronic records through the HIE. The latter option, however, is only possible for Bob’s previous hospitals that are members of the HIE. For those hospitals (or clinics, or physicians) that are not, Alice still must fall back to the paper system.

I explained in section 2 how Alice is likely ambivalent about her preferred format. Of course, there are some doctors who would take great care to request electronic records whenever it was possible. These doctors help to push the adoption of the HIEs. Alice, however, prefers a simpler process for requesting records and doesn’t mind reading through a PDF summary for
new patients. Alice, therefore, has no incentive to use the HIE, as it only complicates her workflow.

This is exactly the problem that the PatientBank team identified at Yale-New Haven Hospital. Even when requesting records from affiliated systems, Yale-New Haven staff will default to sending and receiving faxes, because it is guaranteed to work.

5.2 The Solution: A Bottom-Up Network

After identifying the top-down problem felt by Yale-New Haven, PatientBank offered its record-gathering product to intake staff there. We allow them to externalize the “record requests” by simply specifying a patient, a provider, and a document type. The intake staff, whose jobs were previously operating fax machines, much prefer the new workflow: since we launched at a single department in September, we have expanded to four different departments, with plans to roll out to four more by February.

While it is true that we only gather paper records, we are standardizing the process for ordering them. This is the first crucial piece of building an exchange network. Intake staff will no longer have to worry about request specifics. Rather, PatientBank is in charge of finding the most suitable data with which to fulfill the request.

This leaves us with a strong incentive to gather digital records from other institutions. Over the next few years, it will become easier to request C-CDA documents as hospitals catch up with HITECH regulation. PatientBank, like Rick, cares about data quality, and we will make sure to always gather the highest-integrity record. If we can gather even one electronic document from Bob’s providers, then PatientBank has improved its own data quality. Alice, meanwhile, is satisfied with whichever format she receives via PatientBank.

In addition to asking for digital formats, PatientBank actually facilitates its own electronic exchange mechanism, called **Fulfill**, which allows outside institutions to release documents (whether electronic or paper) by uploading files instead of faxing them. Every PatientBank record request is sent with credentials, allowing the recipient to log in and upload data directly into
the patient’s record. We’ve seen this feature used to upload a C-CDA document, and we view it as promising push towards an all-electronic exchange platform.

### 5.3 Legal Challenges and Conclusion

This paper has outlined technical problems and solutions regarding health information exchange. However, there are many nontechnical challenges when implementing a health record bank, especially legal ones. As noted in section 4, institutions are often required to sign a data agreement before releasing any electronic data. This requirement was set forth by HIPAA, a US law passed in 1996, which put strict limits on who could view and share a patient record without permission. While HIPAA has undoubtedly prevented unauthorized sharing of records, hospitals have been known to abuse it. So-called “data blocking” is the practice of preventing the flow of information, in order to keep patient data inside of private health care networks. This behavior was outlined in a 2015 report to Congress, the “Report on Health Information Blocking”.\(^{14}\) Data blocking is an example of a significant legal challenge, and it is unlikely that technology alone can overcome it.

While legal progress is certainly required, the industry also needs large technical and infrastructural changes to allow electronic data to flow freely. PatientBank’s bottom-up network will help ease the transition, and hopefully bring an end to the days of paper records.

### References


LOINC. URL: http://loinc.org.


Yale-New Haven Health System. Pages - FAQs. ProjectEpic@YNHH.org. URL: https://projectepic.ynhh.org/Pages/FAQs.aspx.