CONNECTING FOR HEALTH℠
MARKLE FOUNDATION A Public-Private Collaborative

LINKING HEALTH CARE INFORMATION: PROPOSED METHODS FOR IMPROVING CARE AND PROTECTING PRIVACY

Working Group on Accurately Linking Information for Health Care Quality and Safety

February 2005

MARKLE FOUNDATION

THE ROBERT WOOD JOHNSON FOUNDATION®

Members of the Working Group on Accurately Linking Information for Healthcare Quality and Safety, Connecting for Health

Clay Shirky, (Chair)
Adjunct Professor, New York University Interactive Telecommunications Program

Ben Reis, PhD, (Staff)
Formerly Manager, Health Program, Markle Foundation

R. Steven Adams
President, CEO, and Founder, Reach My Doctor

David W. Bates, MD, MSC
Medical Director, Clinical and Quality Analysis, Partners HealthCare, Inc.; Professor of Medicine, Harvard Medical School

William Braithwaite, MD, PhD
Health Information Policy Consultant

Jim Dempsey
Executive Director, Center for Democracy and Technology

Daniel Emig
Director, Technology Marketing, Siemens Medical Systems

Lorraine Fernandes, RHIA
Senior Vice President, Initiate Systems Healthcare Practice

Mike Fitzsmaurice
Senior Science Advisor for Information Technology, Agency for Healthcare Research and Quality

Paul Friedrichs
PKI Chief Engineer, Defense Information Systems Agency

Janlori Goldman
Research Scholar, Center on Medicine as a Profession, Columbia College of Physicians & Surgeons; formerly Director, Health Privacy Project

Gail Graham, RHIA
Director, Department of Veterans Affairs

John Halamka, MD
Chief Information Officer, CareGroup Healthcare System; Chief Information Officer, Harvard Medical School

W. Edward Hammond, PhD
Professor, Community and Family Medicine, Duke University; Past President of AMIA

Jeff Jonas
Founder and Chief Scientist, SRD; Member, Markle Foundation Task Force on National Security in the Information Age

Stephanie Keller-Bottom
Director, Nokia Innovent Ventures

J. Marc Overhage, MD, PhD
Chief Executive Officer, Indiana Health Information Exchange; Associate Professor of Medicine, Regenstrief Institute

Peter Swire, JD
John Glenn Scholar in Public Policy Research, Moritz College of Law, Ohio State University; Formerly, Chief Counselor for Privacy in the U.S. Office of Management and Budget

Paul Tang, MD
Chief Medical Information Officer, Palo Alto Medical Foundation

David Weinberger
Publisher, Journal of the Hyperlinked Organization

Table of Contents

Introduction

This document outlines a strategy for linking patient information across multiple sites of care, developed by the Working Group on Accurately Linking Information for Healthcare Quality and Safety, a part of the Connecting for Health effort sponsored by the Markle Foundation and the Robert Wood Johnson Foundation.

The linking of vital information as patients receive care from a fragmented healthcare system is a problem that has consistently plagued interoperability efforts in healthcare. The privacy, technical, and policy issues involved need to be addressed in order to effectively share information across multiple organizations. Making the information available will help to prevent drug interactions and adverse events, avoid medical errors, and help inform decision making for the patient and clinician. It will also enable the support of public health efforts, improvements in research, better physician and organizational performance and benchmarking, and greater empowerment of patients and families as active participants in their own healthcare, among other benefits.

The linking problem is simple to describe but hard to solve: how does a healthcare professional link a patient with their health files, and how do they know that any two files stored in different places refer to the same person? This problem occurs every time a care provider asks to have a patient's file pulled or updated, and every time a patient moves or changes doctors, visits a new lab or specialist, or falls ill while traveling. At its core the linking problem is one of identity -- how can we say for sure that a patient in the office is to be matched with a particular set of records, or that two sets of records can be merged because they belong to the same patient?

The goal of the Linking Working Group was to address these issues, proposing practical strategies for improving healthcare through improved linking of information in a secure and efficient manner, and in a way that allows healthcare professionals much improved access to needed information while respecting patients' privacy rights. Additionally, we assumed that our proposals would be implemented in a five-year time frame, with the additional assumption that any test bed or pilot project implementations would therefore have to be ready in between one and three years, depending on the complexity of the problems to be worked on. We thus focused on techniques for record linking already in use in other areas, rather than on the design of entirely new methods.

Solving the linking problem is only part of the effort needed to improve the healthcare system's use of information technology (IT), of course. There is considerable work to be done on the format and use of Electronic Health records; on the use of available data to improve both medical research and public health;

on the economic models around sustainable deployment and upkeep of these new technologies; and many other issues. Connecting for Health has addressed the broad spectrum of these issues in its “Roadmap” report: **Achieving Electronic Connectivity in Healthcare: A Preliminary Roadmap from the Nation's Public and Private-Sector Healthcare Leaders**, describing in overview a broad vision for improving healthcare through the use of IT. In addition, two Working Groups operating in parallel to the Linking Working Group issued reports on sharing electronic information with patients (**Connecting Americans to Their Healthcare**) and on the business and organizational issues of community-based information exchange (**Financial, Legal and Organizational Approaches to Achieving Electronic Connectivity in Healthcare)**, respectively. These reports are available at http://www.connectingforhealth.org/, as is the response Connecting for Health prepared in collaboration with twelve other influential groups to the federal government’s RFI on the “National Health Information Network.”

The work of the Linking Working Group is meant to address a set of problems that touches almost everyone in the US healthcare system, from individual clinicians to large insurance firms; from local clinics to national hospital chains; from neighborhood pharmacies to state and national public health departments; and so on. Because of this breadth, it has been difficult to find one term that adequately reflects the diversity in size and mission of all the different participants. We have settled on the generic phrases “institutions and providers” or, alternatively “entities” or “organizations” when we mean all participants in the healthcare system, regardless of size, mission, or sector (profit, non-profit, government). As noted in the report below, we also include patients in the list of authorized entities, as we believe the system as proposed will greatly improve their access to their own health information.

In our work, we focused on the problem of linking patient information where the information is widely distributed, and on some of the architectural requirements for supporting that linking in a way that would allow authorized entities to access patient records remotely and securely. Though solving the linking problem would not be a panacea, it would represent significant progress on an issue that is both important in and of itself, and a necessary precursor to tackling other, more complex issues.

Current solutions to the linking problem tend to be ad hoc, paper based, local, and ineffective. Though every institution in the healthcare system from sole practitioners to giant hospital chains faces the linking problem, there is no standard solution, and for many sites of care, paper records are still the norm. Paper records have the advantages of tangibility, making it possible to aggregate individual files easily within a single institution, but are hard to search and hard to share.

As a result, the only files on a patient that can be easily called up are those held locally. Healthcare personnel are thus often forced to work with a very partial subset of the available information on a patient in their care, and frequently end up re-running tests because earlier results are unavailable. At best this creates enormous waste and additional expense. (One of the participants in the Linking Working Group reports that an audit of expenses across systems in Massachusetts found that 15% of expenses were in running duplicate tests because the early results were unobtainable.) At worst, it delays critical diagnosis or exposes patients to invasive procedures unnecessarily.

Background

While the benefits of improved exchange of healthcare information are well known, more efforts to achieve it have failed than succeeded. Problems have included concerns about information ownership, privacy (particularly on a national scale), lack of trust among the participants, the lack of electronic systems in providers' organizations, and the lack of standards that are effective beyond the scope of a single organization.

The Linking Working Group's proposal balances the need to protect privacy with improved discovery and delivery of patient's medical records when they are needed, where they are needed, and only by authorized individuals who need them. The ability to locate patient records and deliver them securely will enable a number of improvements in healthcare, including especially:

- Increasing the ability of authorized clinicians to access vital patient records in near real time in the event of an emergency
- Improving patient access to their own records, allowing them to see and correct mistakes
- Decreasing the number of tests that need to be re-run because the original results can't be found on a timely basis
- Lowering the risk of negative drug interactions because physicians don't know a patient's current conditions or medications

A Decentralized Approach

In approaching this problem we have tried to learn from earlier efforts, but we are also optimistic that the present opportunity offers us a significant advantage unavailable to previous work. Past attempts to create new infrastructure at national scale forced all-or-nothing choices. Often this was because the only models we had for such systems were highly centralized and controlled by the government, e.g. the FAA flight control system or the IRS database.

In the last 5 years, however, we have seen the growth of large-scale but decentralized architectures, everything from AOL's instant messaging system, used by millions daily, to collaborative tools like Groove, approved for secure field use by the Department of Defense. These decentralized architectures mix a high degree of local autonomy with enough global coordination to ensure the functioning of the system. We believe that the flexibility of decentralized architectures offers a way out of the all-or-nothing deadlock. Though much of the required architecture is out of scope for the narrower question of linking patient records, we believe that the architectural characteristics of our proposed approach for linking information is compatible with the needs of a larger health IT system.

The decentralized approach also leaves clinical information in the hands of the clinicians and institutions that have a direct relationship with the patient, rather than moving or replicating it to giant central servers. This approach maximizes the value of incremental development, as the information is already where it needs to be for the system to work. It greatly reduces the risk of misuse, by ensuring that there is no single "bucket" holding clinical information. This decentralization also leaves judgments about who should and should not see patient information in the hands of the patient and the clinicians and institutions that are directly responsible for the patient's care.

It is obviously disappointing not to be able to suggest a scenario in which the whole healthcare system suddenly progresses at all sites, but during our work we have come to the conclusion that such a "Big Bang" scenario actually postpones advantages that can be gotten by working incrementally and on several fronts.

Both a Big Bang and incremental scenario will require significant investment over the next decade, as the healthcare system shifts to more automated ways of delivering information relevant to care. However, any Big Bang scenario would require the standardization of record format, storage, access and transport at hundreds of thousands of sites throughout the US *prior* to the launch of the system.

This would delay by years the value that can be gotten by simpler but more partial upgrades. So long as there is a clear upgrade path and well-defined standards on each of those fronts, there will be steady improvement from even partial improvements in record linking.

Creating the infrastructure for improved linking of records requires some the deployment of additional hardware and software, most of it for the envisioned Record Locator Service and Certification Authorities (detailed below). The clinical records themselves will remain in the hands of the organizations responsible for them. Thus, as with the growth of the fax network or the Internet, the bulk of the IT implementation can be undertaken locally, one institution at a time, and in response to their own needs, budgets, and timelines.

Our work is not complete, having gotten only to the stage where it is good enough to criticize. We continue to work on it within the framework of Connecting for Health, and to present it to knowledgeable people in the health and IT industries, and it will undergo considerable additions and modifications during those conversations. Even early trials of a proposed Record Locator Service (described below) and attendant standards and practices will alter the problems it sets out to solve. As a result, this recommendation is a set of principles, goals, and proposed early tests, but will require constant monitoring and course correction to become effective in any large-scale system.

Working Process

Our goal from the outset was to define ways in which the US healthcare system could be significantly improved over the next five years, through an increased ability to match patients with their existing health records, and through the timely delivery of those records to sites of care.

The Linking Working Group began our process with an articulation of principles that we have used to guide our work. As always with such principles, there is no guarantee that they will not clash, and indeed, there are several such clashes present here. There is, for example, a tension between technological improvement and backwards compatibility. The most backwards-compatible system possible would change nothing, whereas the most radical set of improvements imaginable would require an immediate wholesale upgrade, both distinctly impractical options.

Nevertheless, given the opacity and complexity of some of the issues we are tackling, we have found these principles useful as guiding lights.

Our most basic principles could be combined into a single statement:

> Any proposed solution must support the accurate, timely, and secure handling and sharing of patient records. It must increase the quality of care, the economic sustainability of the healthcare system, and preserve the privacy of patient information. And it must create value for many different kinds of participants, from private, non-profit and government institutions to the individual healthcare professionals and patients who use it.

While that risks sounding like ‘motherhood and apple pie,’ it actually contains several important mutual constraints. We cannot simply trade patient privacy for increased efficiency, for example, or saddle individual providers with unfunded mandates as a way of deploying new tools or technologies.

Beyond this basic statement, we have several additional principles that we believe to be fundamental.

Privacy

Preserving privacy is important to ensuring acceptance of the system and its benefits. Trust is a crucial component of the doctor-patient relationship, including those elements of the relationship that involve the disclosure and sharing of sensitive information. Privacy is an important factor contributing to that trust.

Privacy advocates have long agreed that patients should be informed by providers of the benefits of linking records. However, even well informed patients are reluctant to share information because of privacy concerns. A 1999 survey by the California HealthCare Foundation showed that even when people understood the huge health advantages that could result from linking their health records, a majority believed that the risks of lost privacy and discrimination outweighed the benefits.

The architecture proposed to support the linking of health records has been designed to eliminate the two largest perceived privacy threats associated with the linking of health records: centralization and the use of a unique national health ID. Our approach leaves records with the healthcare providers who created them and uses a person's ordinary name and common identifiers such as date of birth and address to link those records. The only thing centralized is a directory of providers holding patient records and pointers to those files.

In the solution we propose, sharing is peer to peer among participating entities, and both the decision to link and the decision to share records are made locally, where the records are created. The system allows for anonymity and pseudonymity, if agreed upon between the patient and the healthcare provider at the point of service. It incorporates technological advances in authentication, to prevent unauthorized access, and audit trails, to deter and detect abuse by insiders. All of this can occur within the framework of the privacy rules established by HIPAA.

Availability of information

Privacy is only an issue because clinicians must share patient information to do their jobs. Knowledge of existing medical conditions, drug lists, allergies, and other kinds of information about the patient can mean the difference between good and mediocre care and, in extreme cases, between life and death.

And yet, in the US system as it exists today, the main locus for relevant information is not the doctors or labs who have previously seen the patient, but the patient herself. Patients are often asked to remember details about their medical histories, current problems, prescriptions, and allergies, a task they often fail to fulfill. In addition, a provider seeing a patient who has had a test run elsewhere is likelier to choose to have it run again, or to proceed without the results, than to undertake the often futile effort to retrieve the results in a timely manner. (In Massachusetts alone, for example, 15% of medical expenditures have been attributed to redundant testing, costing $4.5 billion per year.)

Increasing the availability of information to authorized caregivers in a secure, accurate and timely fashion is essential to improve clinical care and provide a host of other benefits to the patient and the health system.

Local control of records

Under the system we propose, decisions about linking and sharing are made by the participating institutions and providers at the edges of the network. The system supports (1) linking of records via a directory of pointers and sharing among healthcare providers participating in the system, but it also allows (2) linking without sharing or sharing pursuant only to higher authorization as well as (3) treatment situations that do not result in linking, such as drug or alcohol rehabilitation.

This approach is based on the proposition that we should leave it to providers to determine locally with their patients what to link and what to disclose. By leaving these decisions at the edges, the architecture supports a range of approaches. The default in most systems will be to link and share, the default in others may be to link but not share. The system allows either approach. It also allows higher levels of approval to be set locally for sharing some records.

Patient access to records

A key element of improving linking is making good on the promise of patient access to her own records. The benefits of such access are obvious, from better-informed patients to the possible correction of mistakes and omissions.

There are a number of anecdotes and studies that make a promising case for why it is important for patients to have their own medical records. During the course of our work, one of our Working Group members who runs a system within CareGroup that enables patients to access their records reported a patient catching an incorrect diagnosis of a growing tumor. The patient was able to do this because she realized that the "growth" of the tumor was an artifact of an earlier and incorrect recording of its size. The patient's knowledge of her own medical record saved her an invasive and unnecessary surgical procedure and potentially harmful chemotherapeutic intervention. (See the story of Jerilyn Heinold in ***Achieving Electronic Connectivity in Healthcare: A Preliminary Roadmap from the Nation's Public and Private-Sector Healthcare Leaders***, at www.connectingforhealth.org.)

There are two types of data a patient can and should be given access to -- the audit trail, detailing when and who looked up the location of their records; and the location of their clinical information itself.

There is obvious appeal in having a simple electronic portal allowing for direct patient access. However, the inability to authenticate users securely often prevents implementation of this idea. Current methods used for electronic commerce, for example, use credit cards as proof of authorization, an obvious impossibility here, both because it would lock out patients that cannot or don't want to use their credit cards in this way, and because credit card companies often deal with fraud or identity theft after the fact, an unacceptable option where health records are involved. From studying systems that offer patients access to their medical records, we found that most often first use is authenticated by their provider. Because there is no way to positively authenticate patients remotely during their first use of the system, any patient access must first be authorized, whether in-person or by signature (physical or electronic) by an authorized institutional user.

Once these access credentials are provided, our proposed Record Locator Service will offer a patient remote access (in practice, secure login from a Web browser) to the audit record held in the Record Locator Service, and will provide the same contact and retrieval information to his or her records that the institutions and providers receive. (The patient, of course, will not be able to see any other records but his or her own.)

The patient should be forced to re-authorize periodically, possibly changing passwords as a protective measure (though there is some tension between forced updates to passwords and the patients being able to remember them). The patient should also be able to delegate additional access to their records. One option for such delegation should be the creation of a parallel login, so that a caregiver can be given the same degree of access, with their use audited separately.

Once in possession of location of their records, the patient will still have to request them directly from those institutions and providers, but as that function is HIPAA-mandated and will become increasingly popular, the need to deliver such records at low cost will, we believe, be an additional driver for automation.

Rapid Deployment of Any Solution

Our thinking was strongly guided by the goal of creating significant improvement in five years. The US healthcare system is vast and varied, and any upgrade to its capabilities will necessarily proceed at different rates in different environments, as there is no one entity that could manage such a large and uncoordinated group of entities centrally.

As a result, we have consistently looked for solutions that could be rolled out for testing as pilot projects, and where partial implementation would produce some value. This mandate for marked improvement in five years has also led us to be suspicious of solutions that require Big Bang development, where many pieces of the system are upgraded all at once.

Having articulated these principles, we then attacked the problem of improving the linking of health records. Examined as a problem with the actual matching of a patient with their information as the conceptual core, our proposal has four layers:

(1) Authorized linking of the records themselves, by accurately identifying the patient
(2) Defining the architectural assumptions for support of networked matching
(3) Querying for authorized patient records across multiple institutions and providers
(4) Sharing of requested authorized records between institutions and providers

The Problem of Linking

Better linking of patient records is essential to improving healthcare. Healthcare professionals need patient history, problem lists, medication lists, and a host of other information. The division of labor between primary care physicians, specialists, labs, and third-party payers guarantees that patients' records are spread across multiple sites of care. Currently, each institution has a private silo of patient information, and finding information in another organization’s silos is largely a manual process. A provider wanting a patient's existing but remote records must determine which other institutions and providers to contact and then someone at that institution must locate and deliver the same patient’s record. Aggregating information from these disparate sources is further complicated by the John Smith problem -- how does a clinician or technician gather all the records on *this* John Smith, without also gathering all the information on other John Smiths as well?

Any attempt to improve healthcare IT must solve this problem, since giving healthcare professionals access to information about patients in their care is a core function. Furthermore, this is not just a problem of linking records between different institutions and providers. Operators of large healthcare databases recognize that individual patients are listed more than once in the same and in different databases, within the same institution.

Our recommendation for linking patient records is:

1. The system should not require the existence of a national unique health identifier
2. The system should be designed to create the potential advantages of a national unique health identifier without requiring top-down issuance
3. The system should use probabilistic algorithmic matching of commonly available identifiers to link records

These recommendations are really three parts of the same idea -- design a system for linking authorized patient records using existing demographics and identifiers, rather than waiting for the deployment of new health identifier, but without foreclosing the ability to take advantage of new identifiers should they arise.

The system should not rely on the existence of a national Health Identifier

It is easy to believe that the need for a national system for identifying patients is the same as a need for a single national health identifier. The idea of such an identifier is appealing in its simplicity: give everyone a unique number, to be used only for their health records, so that linking two records becomes a matter of comparing the numbers. If they match, the records refer to the same patient. If they don't match, they refer to different patients.

We defined such a health identifier as having six theoretical characteristics:

- **Unique** Only one person has a particular identifier
- **Non-disclosing** The identifier discloses no personal information
- **Permanent** The identifier will never be re-used
- **Ubiquitous** Everyone has an identifier
- **Canonical** Each person in the system has only one identifier
- **Invariable** A person’s identifier won’t change over time

We then examined the advantages and disadvantages of trying to deploy a national system whose identifiers have these characteristics, and concluded that the disadvantages outweigh the advantages, in part because of the difficulties of designing and implementing such a system, and in part because of the existence of attractive alternative solutions to the linking problem.

We do not recommend waiting for the deployment of a national health identifier to improve the linking of patient information for several reasons. First, it may be impossible to deploy such a system in the United States. Second, even if it were possible, deploying such a system would involve politically complex and sensitive issues that will invariably delay and possibly derail implementation. Third, the imagined end state of such a system presupposes successful solutions to other significant and presently unsolved technical challenges. Finally, the expenses of such a system would be frontloaded, but the value postponed for years.

1. *It may be impossible to deploy such a system in the United States*

Political resistance to any form of national identifier has always run high in the United States, and earlier attempts to discuss the creation and maintenance of such an identifier by the federal government (as no other body would be able to do so) have always been shelved.

Because there are so few cases where proposed national identifiers have ever come close to practical implementation, it is difficult to use past examples when trying to predict the result of any given effort. What we can predict, even in a climate driving increased government inspection of personal data for national security, is that any proposal for a national health identifier will generate violent opposition. The record of success for linking information in the face of such opposition, even for efforts that don't require a common identifier such as Computer Assisted Passenger Prescreening System (CAPPS II), is extremely poor.

Past history and current resistance to government-managed identifiers suggest a high chance of outright failure in any attempt to create a universal identifier for healthcare.

2. Deploying such a system would involve politically complex and sensitive issues

Part of the appeal of a national health identifier is that the existence of such a system presumes that a number of intermediate challenges have been solved. In looking at the existing literature on Health IDs, however, it is clear that many of these intermediate challenges are themselves quite complex. ASTM recently released a proposal discussing a Universal Healthcare Identifier (E 1714), but even at this speculative level, getting a UHID triggers enormously thorny political questions. The report suggests limiting the issuance of such numbers to the US population, but that leaves the question of defining membership in that population. Is the identifier to be limited only to citizens? Does it include Green Card holders? Visa holders? Everyone? These questions are part of what has distorted the use of the Social Security number so badly. For example, the pressure to present such a number to receive healthcare has led many immigrant communities to share their Social Security numbers with one another.

Similarly, much rule making on healthcare is done at the state level, and federal rules that pre-empt state authority are therefore quite contentious. Because of the single-issuer nature of any federally-run health identifier system, it is unlikely that partial or trial implementations will be allowed before these issues and others are settled, making progress on this issue vulnerable to long and possibly paralyzing debate on a variety of topics from immigration policy and the nature of citizenship to the relationship of the federal and state governments.

3. Such a system presupposes successful solutions to significant technical challenges

Decomposing a system for issuing national health identifiers into smaller sub-problems reveals a sizeable number of technical challenges, some of which have never been adequately solved. The system requires the creation of a national system for brokering trusted access to encryption keys, itself a tremendously complex problem. Effective management of public-key infrastructure (PKI) for encryption keys has been an elusive goal for over a decade.

Even the implementation details raise serious barriers to completion: Since no such identifier exists today, every health database in existence would have to be re-worked to store and retrieve patient records using this number. Discussions of a national health identifier system often focus on how it will work when it is finished, but the task of building such a system requires solving difficult sub-problems as well as implementing significant updates across a poorly coordinated system. In any system whose goal is improvement in a five-year timeframe, the technical challenges alone risk pushing such a system out of the realm of practicality.

4. *The expenses of such a system would be frontloaded, but the value postponed for years*

The ASTM Report puts it starkly: "To gain the benefits from such an identifier, it must be used by all relevant organizations." In practice, this means it must be deployed to a significant portion of the clinics, labs, insurance agencies, hospitals and other participants in the US healthcare system before it begins to create any great value. Given that it will necessarily be capitalized by these individual organizations (no one organization, not even the government, could underwrite the necessary technology upgrade), there will be significant inertia to overcome, as everyone will want to postpone implementing a system that will only be really valuable once everyone else has implemented it as well.

Furthermore, during the time the system is being deployed, many individuals in the system will have some of their records tagged with their Health ID, and other records not. That means that during the migration to the new system, there will have to be a method of linking user records among institutions and providers *without recourse to a national health identifier*, the very problem the health identifier is designed to solve.

Furthermore, this requirement is not merely transitional -- since the health identifier is not magic, it would suffer from the problems of all such identifiers, such as accidental transposition, missing fields or other mis-entry of data. (In our surveys of current database practices, one institution reports that even gender is recorded incorrectly in ~3% of cases.) All large-scale systems we examined use multiple identifying characteristics to certify a match, in order to limit the damage any erroneous single identifier could cause. Not only are the match rates of such systems higher than matches using any single identifier, they help uncover bad data as a side effect. Thus the requirement for being able to link patient records without using the health identifier will be a permanent need, in order to handle situations where such identifiers are mis- or un-recorded.

These issues have led us to the following conclusion: Any effort to produce a health identifier will require significant effort and investment; will suffer from a high risk of failure; and will not produce partial improvements when partially implemented. In addition, there will be a persistent requirement for a system that can link user records without recourse to such an ID, even if such an ID were deployed. This combination of high cost simply to secure political agreement, long lead time and enormous expense to get such agreement, and the uncertainty that such agreement could ever be secured, makes us skeptical that work on a national health identifier is the best use of time and resources dedicated to improving healthcare through the use of IT.

Therefore, we believe that the effort and expense trying to make a national health identifier a reality we could better spend on improving the current systems that link patient records using existing identifiers. Furthermore, should a national health identifier or indeed any broadly accepted identifiers come into being, they can be used as additional sources of likelihood of match. No system will ever rely on a single identifier, as some secondary set of information will be needed to resolve ambiguous matches, and any data that can be used for such disambiguation can thus be integrated into the system we propose. Armed with this conclusion, we then set about examining the alternatives.

The system should concentrate on gaining the potential advantages of a health identifier without requiring top-down issuance

We believe we can get many of the advantages with few of the disadvantages of a theoretical health identifier system if we treat participating institutions and providers as globally unique issuers of locally unique identifiers, and if we assume that an individual can have more than one such identifier. This approach avoids the need for a single top-down issuer of identities (and the attendant political opposition and expense), while also allowing the system to grow incrementally rather than requiring Big Bang development.

In practice, an identifier for patient records in any given system will be a concatenation of an identifier for a holding institution and a local record identifier for that patient contained in the master patient index (MPI), creating a globally unique identifier (GUID) for a patient's records in one particular institution of the imaginary form Record-94385723@General-Hospital.

This type of identifier could provide the most critical four of the six possible characteristics of the health identifier, with caveats. Such an identifier would be:

- **Unique** Only one person has a particular identifier
- **Permanent** The identifier will never be re-used
- **Non-disclosing** The identifier discloses no personal information
- **Ubiquitous** Everyone with healthcare information has an identifier

Uniqueness and permanence can be provided so long as HIPAA-mandated identifiers of healthcare organizations and providers are themselves unique, and so long as an institution does not re-use in-house patient identifiers.
(A "No re-use of identifiers" policy is widely regarded as best practice in database management, but is not universally adopted at present.)

Such an identifier will be non-disclosing if the institution uses a numerical string or other non-semantic pointer to a patient. (This will require participants not to use the Social Security number as the patient identifier. Though the risk of identity theft is already driving this change, the non-disclosing requirement would make it mandatory.)

Ubiquity is definitional – since being listed in the system required the presence of such an identifier, the identifier will be (tautologically) ubiquitous. The larger challenge is to extend the system to the broadest possible adoption in the shortest time. Another requirement imposed by this approach is a unique identifier for the healthcare organization and provider, which HIPAA has mandated by 2007. (Pilot projects and other early work will require the issuing of temporary versions of such identifiers.)

A comparison with e-mail is instructive here – there are many John Smiths, but only one John.Smith@IBM.com. IBM.com is a globally unique entity, and is responsible for the local uniqueness of email addresses in its domain. Likewise, John Smith might also have js1964@gmail.com, also a valid and globally unique email address. In such a system, a person will have several such pointers to medical records, as they do today for their records that exist in multiple places, and there is no guarantee that any one identifier will point to a patient for life.

Thus the two characteristics of a health identifier this system forgoes are canonicalness and invariability. Though these would be desirable characteristics if they could be obtained at little implementation cost, they are not requirements for successful identity systems (as the email example shows), and they are the two characteristics that that create the greatest difficulty in implementation, and create the requirement for a single issuer of identity (in practice, the federal government, whether directly or by proxy). We believe that a system without canonicalness or invariability is better suited to incremental creation of value and to shared participation among a large number of otherwise uncoordinated actors.

The system should use algorithmic matching of commonly available identifiers to link records

Assuming that patient records can be globally identified using the Record-94385723@General-Hospital format, the challenge in such a system is to link a patient's records across space (linking records in different institutions and providers) and time (finding historical records associated with the patient sitting in your office today). This problem occurs today, whenever two institutions or providers need to share information on a patient -- primary care physician and specialist, clinic and lab, hospital and HMO. Our proposed solution for handling these cases is probabilistic matching of the patient, using existing patient identifiers.

You can think of this technique as progressive exclusion of non-matching records. Imagine a doctor refers a John Smith to a hospital, and the hospital staff wants to pull out Mr. Smith's medical history from their files. The hospital would start by excluding all patients in its database whose last name was not Smith, then, for this population, excluding all Smiths whose first name was not John (or variants, such as Jack), then excluding all John Smiths whose birthday was not March 4, 1954. These stages of exclusion would continue using all available identifying data -- address, phone, social security number, and so on. (In practice, of course, the process is automated and instantaneous. It is described in stages here for illustration.)

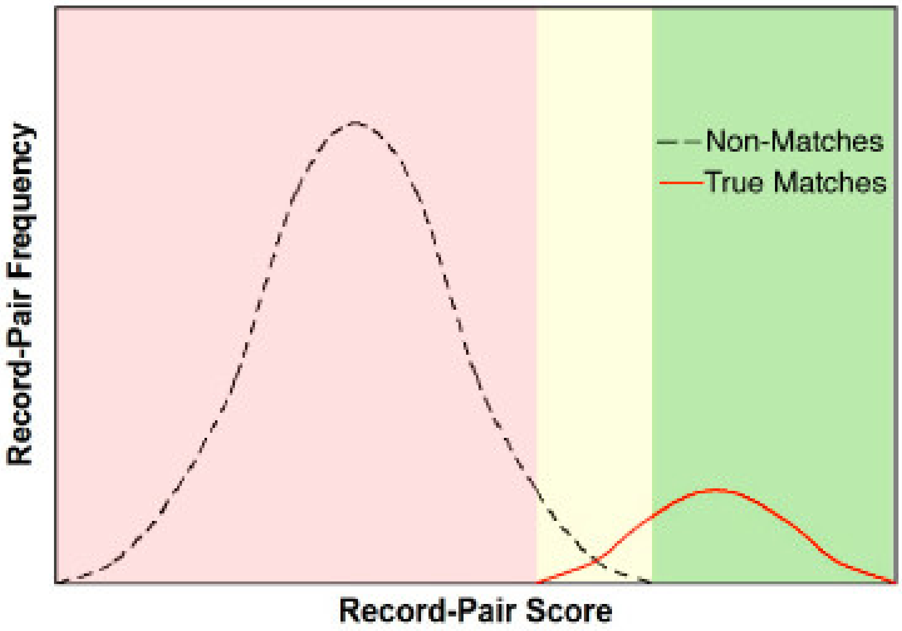
Such a system can operate with a single cutoff for a match (e.g. "Treat these two records as belonging to the same patient if first name, gender, date of birth and SSN all match"), but the system can be further improved by weighing the probability that similar information in different records indicates that the records belong to the same patient.

Probability-weighted matching can improve the quality of record matching by taking the specific characteristics of records in particular databases into account. In a theoretical database where last name was similar in 90% of accurate matches (there were some misspellings), while it matched accidentally in 2% of cases (accidental overlap), a matching last name would be 45 times more likely (90 divided by 2) to be a feature of an accurate than false match. Each available field will have a characteristic prediction of match -- no one field is perfect (and, so long as humans enter the data, no one field can ever be perfect), but the ability to assert an accurate or false match grows with every field compared.

Possible Categories of Record Pairs

The smallest possible unit for considering the linking problem is a pair of records, each of which contains some identifying characteristics (name, gender, DOB, SSN, etc.). These records can be characterized in two ways. The first is whether they *seem* to match, which is to say whether the identifying details in the two records are similar enough to indicate a possible match. The second way of characterizing a pair of records is whether they *should* match, which is to say whether they actually refer to the same individual.

You can give record pairs a 'potential match' score -- low to high likelihood of a match, and record pair frequency -- number of records that have a particular score. Such a graph will have this rough distribution, where the area under the dotted line contains record pairs that do not refer to the same individual, while the smaller area under the red line contains record pairs that do:



The most common category by far is obvious non-matches, in the shaded area on the left. These are low scoring record pairs that do not refer to the same person. A record for Susan Smith, DOB 3/9/1969 is not to be linked to a record for Anthony Moon, DOB 4/5/1997. The most important category is high-scoring pairs where both records actually refer to the same person, the area shaded here on the right. Improving the ability to find and make such records accessible is a core goal of improving linking.

The greatest challenge lies in the middle region, which contains the records whose scores are high enough to be possible matches, but not so high as to be obvious matches. In an ideal world, we would be able to completely separate non-matches from true matches, but because of the overlap, the middle zone contains two additional categories. Record pairs that have a high match score but do not actually refer to the same person are false positives, while record pairs that have a low score but do refer to the same person are false negatives.

Thus the interaction between record pairs that have a high match score and those that actually refer to the same person creates four conceptual categories:

High Score Same Patient	Yes	No
Yes	**True Match**	**False Negative**
No	**False Positive**	**Non-match**

Despite the similarities, the outcomes of false negatives vs. false positives in a clinical setting are radically different. The US healthcare system currently functions under the assumption of prevalent false negatives -- caregivers are accustomed to operating with incomplete patient information. False negatives, while undesirable, are normal and, in any system that protects patient privacy, also inevitable. No system that allows patients to opt to keep certain records out of view can also guarantee that caregivers have a complete clinical record. Thus a goal of improved record linkage is to greatly enhance access to relevant information, without ever pretending to guarantee 100% coverage of all of a patient's records.

False positives, on the other hand, can be catastrophic, as they can lead a caregiver to wrongly believe they have information that may have life or death consequences. A doctor given incorrect medication or allergy lists, for example, may prescribe an inappropriate drug resulting in significant and negative consequences, where patient records are inappropriately disclosed by being incorrectly combined with the records of patients with similar names. Thus the first critical design step in pulling records on a particular patient is to raise a very high threshold for matching data, in order to optimize the system against false positives. This will of necessity raise the number of false negatives, but this is a distinctly less bad outcome than allowing false positives and false negatives to appear at similar rates.

Once this high threshold for a presumed match has been created, the second, critical step is to use the available identifying characteristics to remove the remaining false positives, leaving only true matches. This is the function of probability-weighted matching algorithms.

Construction of Probability-Weighted Matching

The process of probability-weighted matching works by taking two groups of records, whether in different institutions or providers or from different databases in a single institution, and comparing each record in database A against each record in database B, thus generating a sample population of record pairs. Most of these will be non-linked pairs, which do not belong to the same patient. In turn, most non-links will have little or no overlap in identifying characteristics, because the records belong to different patients, and thus contain significantly dissimilar identifying characteristics.

However, some pairs will score high enough to indicate that both records may refer to the same patient. These pairs of records will contain true matches, which should be linked, as well as non-matches, pairs that seem to belong to the same patient, but don't. Asserting that non-matching records refer to the same patient produces a false positive match. The goal of probabilistic matching is to pull out as many true matches as possible, while producing few false positives, ideally zero.

To do this, the corresponding fields of every linked record pair are compared, to see how likely it is that a particular field, such as last name, matches in similar vs. dissimilar records. The important calculation is the ratio of correct data in true matches vs. incorrect data in non-matching records overall. Some initial predictor for asserting a true match needs to be produced, but simply acts as a stake in the ground for further refinement of the measurements. For the purposes of the discussion below, we use data drawn from the 2002 paper *Analysis of Identifier Performance using a Deterministic Linkage Algorithm*, by Shaun J. Grannis MD, J. Marc Overhage MD PhD, and Clement J. McDonald MD. (Marc Overhage was a member of our Working Group, and Shaun Grannis provided comment on a draft version of this document.)

Performance of Probability-Weighted Matching

In that study, the authors compared a group of records between institutions A and B, using a match on SSN as the beginning predictor of a match. True matches without common SSNs were left as false negatives; the principal work of the study was to separate low-scoring true matches from false positives.

In this patient population, last name was the same in 93.5% of true matches (the last 6.5% being accounted for by last name change, data entry error, etc.) In the same population, where SSN matched, last name was the same in 21.6% of non-matching pairs. Thus, a matched last name is 4.3 times more likely to occur in true matches than in non-matches. (The unusually high co-valence of matching last names is an artifact of using SSN as the predictor of a link, which tends to match among people with the same last name. Multi-variate predictors of matches will have fewer artifacts of this sort.)

You could perform this calculation for every possible identifying field. Gender, for instance, has a better chance being correctly recorded and unchanging than last name does, being accurate in roughly 97% of cases. However, gender also overlaps by chance in roughly 50% of cases. Thus, though gender is usually more accurately recorded than last name, it is only a little less than twice as likely to be the same in true matches as in non-matches.

First names are more complex. There are more first names than last names in the US population, making them better predictors of true matches. Variable spelling (Marcia, Marsha) and the acceptance of nicknames as synonyms (William, Bill) complicates the match prediction problem. The use of name-similarity databases such as Soundex, Metaphone, and the New York State Identification and Intelligence System algorithm (NYSIIS) can greatly increase the predictive value of a first-name match. Matching NYSIIS-transformed first names were found to be present in 89% of true matches and only 1.4% of non-matches, or 63.5 times more likely to occur among true matches than non-matches.

Date of birth exhibits still another pattern, where each of the sub-elements (day, month, year) can be analyzed as a match predictor. The AIMA data indicated that a match on day was 11.5 times more likely to occur in true matches than non-matches, month was 19.4 times more likely, and year was 22.2 times more likely. The advantage of treating date of birth as a collection of sub-fields is that even in the event that one element is missing or incorrectly recorded, there is still some predictive value in the remaining fields.

Multiplicative Value

The principal value of such variables is not in isolation -- no one would try to identify a patient based on a single characteristic, not even SSN -- but in combination. And, critically for the algorithm, the combinations are multiplicative. For example, a complete match on all three DOB fields (day, month, year) is almost 5,000 times more likely in a true match than a non-match in the patient population involved. Similarly, in a population where a similar first name is 63.5 times more likely to refer to a true match, and date of birth is 4,953 times more likely to do so, matching first name and Date of Birth is more than 300,000 times more likely to do so. (This multiplicative effect is variable, however, depending on the fields being concatenated. Ethnicity, for example, means that first names and last names are not completely independent variables. Likewise, first name is strongly correlated with gender, so knowing gender does not double the accuracy of knowing first name.)

The multiplicative value of prediction can be illustrated using patient populations. In the patient population covered by the union of databases A and B, among pairs of records that match by first name and full date of birth, there will be one false positive for every 33,000 true matches, or, put still another way, a better than 99.9997% accuracy rate for true matches over false positives. (False negatives, as noted above, will necessarily occur at a higher rate, but these are a much less undesirable outcome than false positives.)

These rates improve further when secondary characteristics can be matched on, such as SSN or Zip code. Furthermore, matching on multiple variables is robust and can protect against the occurrence of certain non-matching characteristics, such as inaccurately entered data or changed last name. Note that even if a universal health identifier of some sort existed, such a process would be needed in the event of missing or mis-recorded identifiers, and could use such an identifier to improve the matching algorithm, even given the inevitable inaccuracies in at least some of the recording of such an identifier.

The method as described here is a greatly simplified version of a more complex and iterated operation. In particular, pre-processing of the data can produce much more accurate inputs, by grouping sound-alike and nicknames as noted above, or by analyzing numerical data for simple number transpositions, which can increase the predictive rate for fields like Date of Birth, Zip, and SSN. The critical question is which combination of fields will produce such a high likelihood of accuracy that the number of false positives produced will be miniscule compared to the number of true matches recovered.

We have many working examples of multi-field linking being used effectively across multiple databases covering in excess of one million patients. More work is needed to determine the effects of increased scale in the population to the 10 million+ range, and to determine the effects of increased geographic spread. By covering a large area, such a system would lose the predictive advantage of geographic locality (most healthcare is local, so a high-scoring potential match drawn from the same local pool has a higher chance of matching) but gains greater heterogeneity of last names, possibly improving the validity of that field. More work is needed to understand the effects of such changes in scale and scope.

Data Completeness and Cleanliness

Of course, the critical issue is the availability and cleanliness of the relevant data. Large parts of the healthcare industry, and especially much clinical practice, remain heavily dependant on paper, limiting the availability of even the most basic identifying characteristics in electronic form. Furthermore, inaccurate and duplicated data is common, even among institutions with a high degree of automation in handling patient records.

For the short term, any work on pilot projects must make the existence of a patient's identifying data in electronic form a pre-condition for participation. The question of what will lead the myriad small providers who make up much of the healthcare system in this country to upgrade their record handling systems will be beyond the scope of any short-term test.

Longer term, however, work must be done to understand how to provide both the necessary technology and incentives to get providers to collect patient data in an electronic format, and to use this data as part of the linking infrastructure outlined here. In all likelihood, this effort will be linked with other efforts to improve the storage of clinical information in an Electronic Health Record format. In addition, work needs to be done on methods and incentives for improving existing records, both merging duplicate records and updating incorrect fields in existing records.

Given the diffuse incentive structure of the US healthcare system, some sort of pay-for-performance incentive for gathering accurate records and cleaning existing ones is one obvious possibility to explore. Such a change, however, will almost certainly be part of a larger re-alignment of incentives in the direction of use of IT, and thus can't be easily integrated into any narrower test of particular capabilities such as linking patient records.

Real-World Implementations

To work at any large scale (millions of patient records or more), such a system of probabilistic matching must have enough identifying characteristics about the patients to make one-and-only-one matches in the majority of cases, and must produce a negligible the false positive rate.

Because of these requirements, the Linking Working Group was initially skeptical that probabilistic matching could work in a large (and ultimately national) network of linked healthcare providers. However, as we uncovered research in the field such as the work by Grannis, Overhage, and McDonald, as well as examples of healthcare systems that were using such matching while handling millions of records (e.g. Sutter Health, the North Carolina Immunization Registry), and similar systems outside healthcare (Defense personnel, Las Vegas casino staff), we came to the conclusion that such a system is not only workable, but is already working at large scale in many places.

Partly as a result of these early examples, we then decided to survey the linking practices of a number of healthcare institutions who met the following criteria: large patient population, spread out among a number of institutions, thus requiring some form of distributed linking. The institutions we surveyed were CareGroup (Boston, MA), the North Carolina Emergency Department Database (NCEDD), Provider Access to Immunization Registry Securely (PAiRS, also in North Carolina), Regenstrief Institute (Indianapolis, IN), RxHub, Santa Barbara County Care Data Exchange, Santa Barbara, and Sutter Health (California). These interviews confirmed our earlier sense that probabilistic matching can be effective at large scale. (A narrative description of the survey appears at the end of this document.)

On reflection, the discovery of effective multi-database record linking was less surprising than it first appeared. Since these institutions are, by definition, operating without a canonical patient identifier (the Social Security number being hopelessly compromised), and since the multi-party nature of the US healthcare system requires constant collaboration among doctors, hospitals, labs, payers and a host of other participants, there is an enormous incentive to get the patient identifier right using whatever identifying attributes are to hand. Nevertheless, we were heartened to discover how much work, both theoretical and practical, has been done on probabilistic matching.

The existing systems are not perfect, of course. Successful use of such systems requires that the participating entities capture a number of identifying characteristics; that the data be relatively clean; and that there be a minimum of duplicates and data-entry errors. Even when these criteria are met, the system will still generate a number of ambiguous results, requiring either careful performance tuning to make sure that these do not become false positives (for the reasons noted above), or a staff trained to make the judgment calls the machines are incapable of.

These requirements, though, are still less onerous than what would be needed for a national Health ID, which also requires clean data entry and database access, but would also require propagation of an entirely new standard, even to systems that currently meet the other data requirements.

Probability-weighted matching has two other advantages that Big Bang proposals lack. First, the chance of false matches rises only gradually with scale. In a clinic with only hundreds of patients, first+last name alone will be enough to identify most patient records uniquely, so sharing those records among a small number of small clinics will not create the same issues of name clashes as a multi-million record system. Thus small providers can improve incrementally, as they interconnect incrementally, cleaning data and capturing new fields as they grow, rather than re-engineering everything all at once. The second advantage is that large service-oriented systems such as labs and pharmacies are already well along the path towards clean, more queryable data, offering even small providers immediate value for plugging into a network of health data.

Recommendations for Probability-Weighted Matching

We believe the advantages of probability-weighted matching outweigh the disadvantages, and will in any case be required both as a transitional and later as a back-up system, should an alternate solution be adopted. We therefore recommend further development, with particular concentration on the following areas:

1. Document current practices and possible improvements
2. Document current practices in data capture and cleanliness
3. Explore incentives for better capture and cleaning of data
4. Develop a reference implementation of such probabilistic matching, including especially an open format for passing such data over secure network connections
5. Develop a pilot project that tests and improves the reference implementation

Document current practices and possible improvements

Our simple and qualitative survey of large health systems has convinced us that deep knowledge about probabilistic matching exists in many places. An obvious next step would be a more quantitative comparison of the specific algorithms used for matching. Of particular concern in such a survey will be practices around data entry, data cleanliness and the merging of duplicate records and purging of inaccurate records.
We would also want to uncover which auxiliary databases are in use, such as sound-alike and nickname dictionaries, and which additional sources of data are used (e.g. non-traditional identifiers such as Zip+4, mobile phone numbers, etc) to aid in more accurate disambiguation.

We have benefited in the Linking Working Group from representation of members who manage large-scale data matching programs outside healthcare. We would want to survey additional organizations with linking technology or research efforts in other realms, from travel to law enforcement, to profit from external expertise.

Document current practices in data capture and cleanliness

As the system is designed to allow for incremental upgrade, some level-setting needs to be done at the outset, in the form of a survey or other research instrument that helps determine the current state of data capture and cleanliness. There will doubtless be a range of quality, from simple paper-records where the patient's names are the principal identifiers to fully queryable multi-institution databases. A second goal of this research should therefore be to determine and describe the characteristics of the entities with the best data practices.

Explore incentives for better capture and cleaning of data

Even more important than knowing the current state of practice is figuring out how to improve data collection and cleanliness, one institution at a time. The valuable effects that can come from pooling information about a patient can only be attained if the data is clean enough to be worth linking -- a database whose records are too dirty will generate more false negatives than true matches.

It will be critical to understand both the incentives and hurdles to improving the acquisition and handling of data. One advantage in improving networked use of data is that improved accuracy will create entirely local benefits as well, as entities can lower the cost of capturing, storing and handling their own data, even with no external partners involved.

Understanding how to help participants undertake the necessary upgrades and changes to process, ideally out of local interest, will help advance the larger goals of accurate linking.

Develop a reference implementation or prototype, including especially an open format for passing such data over secure network connections

In our early survey of healthcare organizations, we found a number of commonalities; most organizations use Name, Gender, Date of Birth, Address, and Social Security Number as input for matching. However, different organizations put different weights on different fields. The goal of any survey would be to develop a consensus view for whatever practices seem broadly supported, and to adopt or design a set of practices around whatever areas lack industry consensus. This would not require all providers to supply exactly the same fields, but rather to provide as much of a core subset as possible (first name, last name, date of birth, gender), and to offer whatever secondary characteristics they have, such as address, Zip, phone, et cetera.

This set of techniques would then be instantiated as a working system for matching patient identity in order to link records, and the results published as a reference implementation or prototype. Particularly important in this reference implementation will be the design of an interchange format, a document format that will contain the fields parties will be assumed to be using in matching patient records. This format should be designed to allow organizations using different databases and tools to exchange the data necessary for linking without requiring global engineering efforts. Instead, the only thing a participating organization will need to implement is a local translator between their internal format or formats and the exchange format, thus simplifying the interoperability challenges.

Develop a pilot project that tests and improves the reference implementation

The motto of the Internet Engineering Task Force is "Rough consensus and running code." This is an implicit admission that the ability of even experienced and talented engineers to predict how large scale systems will perform is limited, and that gaining practical experience is better than creating a perfect theory.

The reference implementation should therefore be field-tested, in one or more pilot projects, in order to see what actually works and what doesn't. These pilot projects should be designed to test the reference implementation in real-world settings and to document the process of building and running such a system in order to identify practical bottlenecks. The design of these projects will require significant and ongoing attention to the myriad practical details of health IT, and thus cannot be specified completely without significant input from the actual participating organizations. Thus a first goal for launching pilot projects should be the recruitment of organizations who are willing to help design and test the linking techniques included in the reference implementation, and to participate in the design, construction and testing of the additional infrastructure, including especially the Record Locator Service (described below) necessary to make patient linking part of a larger healthcare network.

Of particular interest in the design of these pilot projects will be finding organizations willing to work together outside the typical regional framework for such projects. This can be achieved by working with healthcare providers from different regions, or with organizations that do not have a specific regional focus (e.g. RxHub).

Architectural Principles

Because the linking problem involves multiple organizations and providers, it is necessarily a network architecture problem as well. Though there would be some advantage in improving the ability to link records among different databases internal to an institution, the most complex linking issues appear when patients are moving between different localities or sites of care. As a result, along with improvements in probabilistic matching, the Linking Working Group focused on how different organizations would be able to run such matches on data held elsewhere.

The architectural vision described here is focused primarily on federating the ability to identify a patient's authorized information held remotely. It necessarily touches on subsequent challenges any fully-fledged system will have to address -- locating, sharing, and interpreting that information -- but specifies those operations in less detail.

The Linking Working Group approached the issue of architectural aspects of linking in much the same way that we approached our work as a whole; we began by articulating a set of architectural principles, which we called requirements and constraints. Armed with these principles, we then laid out a high-level overview of what such an architecture would look like, with an eye to supporting the querying for the existence and location of records and the sharing of records once found.

Design Requirements and Constraints

We began our conversation about architectural support for linking by detailing what we believe to be fundamental requirements or constraints on the design of such a system. These design considerations are sometimes technical expressions of our broader principles, listed above; much thought went into architectural support for patient privacy, for example. In addition, some of the design considerations are drawn from the literature on the development of large heterogeneous systems, especially the Internet and the Web.

Though we listed these requirements as inputs to the design of a record-linking system, our architectural discussions became inputs for the Connecting for Health “Roadmap”: (**Achieving Electronic Connectivity in Healthcare: A Preliminary Roadmap from the Nation's Public and Private-Sector Healthcare Leaders**) as well (available from http://www.connectingforhealth.org.) We believe they are useful principles for the development of broader health IT architecture, not just linking.

The core requirements and constraints are:

- Decentralized
- Federated
- Built without requiring 'Rip and Replace'
- Built through decoupled development
- Built on top of the Internet -- no new wires

Decentralized

The US healthcare system is fragmented. Many types of organizations exist as part of the current healthcare network, from giant hospital systems and insurance agencies to individual practices, with all manner of specialists, clinics, and agencies in between.

We are confident in predicting that this situation will still hold true in five years time. Therefore, any proposed improvement to the healthcare system must assume that the participants will be decentralized, and must be designed to accommodate at least some voluntary, partial, and incremental participation.

Federated

As a related principle, the actors in the system must be coordinated to some degree, as with agreements on standards or practices, but cannot be managed in any sort of command-and-control or top-down fashion.

Furthermore, the requirement that each participating entity be able to control the release of the patient information it holds guarantees a high degree of informational sovereignty to local systems.

As a result, we believe the system must be a federation, where a mutual and mutually re-enforcing set of standards and agreements bind individual participants. Federation, in this view, is a response to the organizational difficulties presented by the fact of decentralization.

No 'rip and replace'

As has been noted in our meetings numerous times, one cannot take the healthcare system down for the weekend in order to re-tool it. The strictures of economic sustainability and practicality demand a clear migration path for participants in any health architecture.

In particular, we are skeptical of the rip-and-replace style of development, which assumes all participants will remove their existing systems, thereby destroying the staff practices that go with these systems, and replace them with a uniform set of software, requiring wholesale re-training of staff.

This is not to say that there will be no new standards, software, or practices -- all are essential. It is simply to say that as a result of the decentralization noted above, the adoption of novel features of the system will proceed at different rates in different regions and for different actors, and that where possible, the adopted software, systems, and practices should be incremental to current ones.

Decoupled development

When too many separate standards are bundled together in an all-or-nothing package, the expense and organizational difficulty of upgrading everything all at once becomes prohibitive. In order to allow local organizations the discretion to choose which upgrades to pursue when, we imagine a set of standards that can be implemented in various orders, and can be effective at various levels of completeness. Though the hope is that all organizations will eventually upgrade record location, sharing, and formatting, the order and pace of these various efforts must be informed in part by local conditions and decisions.

The system we recommend is designed to assume minimal and various thresholds for entry into the system, on the assumption that by offering some value in return for some embrace of standards, we will be able to maximize early membership in such networks. Once in, the members will have both the incentive and opportunity to become increasingly standards-compliant, and therefore to have increasingly high levels of interaction with one another.

Built on top of the Internet -- no new wires

As a practical matter, we believe the Internet will be the default method of connecting participants, and of providing connection to services. No new wires should be required to make the system operational. Use of the Internet for health information will require the definition of standards for encryption of data, required security measures, and auditing to allow for accountability and oversight, in order that all participants can trust their partners to preserve patient privacy and data integrity.

Adherence to such standards must be a requirement for use of the system, and violation of such standards will be cause for censure or even, in extreme cases, ejection from the system.

Any proposed change must take into account the current infrastructure of the healthcare system, and must work with that infrastructure where possible. Some of this infrastructure will need to be replaced, of course, and the replacement and migration will generate new costs, if only during the period of transition, but where possible, the system should work alongside what has been deployed today, and the changes, when they come, should ideally be staged so that they can be adopted gradually over time.

Top-down and bottom-up

We regard the debate between a bottom-up approach (many local initiatives) vs. a top-down approach (a single national one) as fruitless. Most healthcare is local, and many multi-institution systems that serve particular localities already exist. However, to ensure interoperability between those regional systems as they grow, some national standards must be in place.

The question is not whether to work from top up or bottom down—both are necessary. The question is which problems are most amenable to which type of solution. How an institution chooses to store and retrieve patient records will be local because it *is* local – there is too much diversity in medical record keeping to impose a single national set of tools and techniques. Instead, we will start by recommending best practices and working towards migrating all players to a common set of supported standards, but will assume that local diversity will continue for the foreseeable future.

Meanwhile, things like minimum security standards for secure Internet transmission or patient matching methods must be national, so that all participating organizations can connect to one another securely and without significant protocol mismatch. Here we will work in a top-down fashion, as the problem demands it. (See a discussion of the "Common Framework" in **Achieving Electronic Connectivity in Healthcare: A Preliminary Roadmap from the Nation's Public and Private-Sector Healthcare Leaders** at www.connectingforhealth.org.)

Ultimately, the design challenge is the federal one: leave to the local systems those things best handled locally, while specifying at a national level those things required as universals, in order to allow for interoperability in those areas where the local systems must communicate or share.

Protecting privacy locally

The question of privacy in such a system is critical, both because of legal requirements and because patient trust in such a system is essential for success. Under the system we propose, privacy decisions are made locally, based on a patient conversation with the healthcare provider. Protections on older information are governed by conversations with previous providers that occurred at the time the relationship was established. In our proposed system, retrieval of records involves a two-step process: First, the requester queries the directory and gets pointers to any authorized records indexed in the directory. Then each provider holding records has the discretion to disclose, depending on that provider's rules, as defined in the provider's initial encounter with the patient. Thus, there are two decisions to be made locally: whether to index and whether to share.

It is easier to protect privacy locally. It is hard to make restrictions travel with information or to centrally keep track of privacy preferences. Lessons drawn from the context of digital rights management show that protections linked to data as it travels can be easily defeated or ignored. Furthermore, the complexity of privacy preferences that could be expressed in medical records would be hard to scale. The vision of mobile but self-protecting medical records is simply not feasible. What is feasible, and what our system allows, is the protection of records at the place where they are created and where they reside. If a provider restricts records as a matter of policy or as a matter of patient choice, that provider need not index the patient in the directory and, if the provider does provide a link, it need not respond to a request to share.

Under the system we propose, anonymity and pseudonymity can be achieved by the patient locally. The patient can use a pseudonym. Indeed, in some situations, pseudonymity may be the norm, as for example, with victims of domestic violence seeking medical treatment. Pseudonymity could even offer the possibility of opting-out on the visit level, if the patient and provider choose to create a pseudonym just for that visit. Or a provider may, as an option, give a patient a pseudonymous identity package to be used over time. Within a particular local system, it may even be possible to link records pseudonymously: the patient may be assigned a single global internal ID when the data goes to the local database, allowing retrieval of pseudonymous data via true name. The GUID would be used only locally, and the broader directory would treat the true name and the pseudonym as separate patients.

Pseudonymity and anonymity may strike some as an unreasonable goal in a healthcare system. However, there are some situations that justify non-disclosure, such as those in which patient or, in many cases, state laws determine that particular medical information is too sensitive to share.

One way to ensure the option of non-disclosure is through anonymity or pseudonymity. Because these instances, though rare, are nevertheless required in some cases, the question of anonymity and pseudonymity has been a stumbling block in earlier linking discussions. The system we propose allows for anonymity or pseudonymity in those instances where it may be desirable.

The system neither guarantees nor forecloses patient anonymity. That is a decision to be made by the patient and provider together; whether a patient's identifiers are reported to the directory is a local decision. In a one-time encounter with a provider, a patient can ask that the records not be indexed in the directory. If the provider complies, the information cannot be located via the Record Locator Service. In the case of an established relationship, a patient can keep records out of the system by asking that they be stored locally under a pseudonym. Even if the records are linked, they cannot be located by the patient's true name. If the patient remembers the pseudonym and associated identifiers, the records can be retrieved in the future.

Close to the local level, systems can set additional rules for access. Higher levels of approval can be set locally for sharing some records, or the system can provide notification to the creator of sensitive records when they are accessed.

The vision that a patient should be able to say, “You can share this record, not that record, this particular piece of information, not that one” is a vision that cannot be easily implemented currently. The complexity of the healthcare system makes it very hard to fulfill this kind of request with high accuracy. A patient's HIV positive status, for example, can be inferred not just from a label on a chart but from problem and complaint lists, medication lists, and written doctor's notes, discharge summaries, imaging studies, etc.

It is our goal to design a linking system that works with the realities of healthcare record systems as they are being designed. Though some domains such as intelligence sharing have sub-record-level permissions for sharing, the healthcare industry typically does not, both because the patient is the key entity in the system, not the individual record, and because health information is still highly unstructured. Most health record systems do not allow for record-by-record distinctions between what can and cannot be shared. It would be a disservice to both patients and professionals to create the expectation of such highly granular and controllable records in today's systems. Such a high degree of granularity is not required by law and is not being implemented in most new systems.

Payment systems are separate and pose their own privacy issues. Indeed, while it is possible to include payment pointers in the directory of healthcare records, so that insurance information is available to providers, it is also desirable to decouple the payment system from the treatment system for day to day transactions yet have the capability to conduct authorized sharing for payment, treatment, and operations.

Under the system we propose, a patient can also collect information in her own home if she wishes, by using the Record Locator Service, then making HIPAA requests from the organizations who hold relevant records. Indeed, the system makes it easier for patients to find and compile their own records than today.

Architectural Overview

With these requirements and constraints listed, we turned our attention to defining an architectural approach to making records available where and as needed, given the linking recommendations above.

The core architectural idea of our proposal is that patient records must remain in the hands of the organizations who create or manage these records – clinics, hospitals, labs – but these records must be readily locatable by other institutions and providers who have responsibility to the patient. Examples include healthcare while traveling, chronically ill patients being treated by multiple clinicians, emergency care, patients changing physicians after a move or healthcare plan change, and so on.

The system will rely on optimizing the current methods of institutional and provider record keeping, using local record numbers of a patient at each site of care, while improving interoperability of existing systems and methods for handling patient records locally. Current work on systems for distributed supercomputing or storage suggests that the problem of interconnecting multiple nodes is best solved using a "connection broker" pattern. A connection broker is a database that maintains records of distributed resources and matches requests with the holders of the appropriate resources.

This central database of pointers can be quite small, relative to the enormity of distributed resources that can be identified through it. Furthermore, for systems operating on the Internet (as we assume this one will), once the organizations involved in information sharing are identified to one another, they can share the requested data directly, without further involvement by the connection broker. (However, in some cases, it may be desirable to set up proxy or caching servers, to allow less technically sophisticated clinics, hospitals and other users access to the system.)

Record Locator Service

Our system imagines the construction of such a connection broker, the Record Locator Service. The Record Locator Service is a new piece of infrastructure. Numerous RLSs would exist in different regional or sub-networks throughout the US. The RLS is subject to privacy and security requirements, and is based on open standards.

- The RLS holds information authorized by the patient about where health information can be found, but not the actual information the records may contain. It thus enables a separation, for reasons of security, privacy, and the preservation of the autonomy of the participating entities, of the

function of locating authorized records from the function of sharing them with authorized users.
- Release of information from one entity to another is subject to authorization requirements between those parties; in certain sensitive treatment situations patients or providers may choose not to share information.
- RLSs are operated by multi-stakeholder collaboratives at each sub-network and are built on the current use of Master Patient Indices.
- The Record Locator Service needs to enable a care professional looking for a specific piece of information (PCP visit or ER record) to find it rapidly. An open design question is how and where in the model this capability can best be accomplished.

(For more on the Record Locator Service and the proposed Standards and Policy Entity which would set the guidelines by which it would operate, see the response Connecting for Health prepared in collaboration with twelve other influential groups to the federal government's RFI on the "National Health Information Network" at www.connectingforhealth.org.)

To achieve these goals, we focused on two functions of the network:

- Finding places where a patient might have information
- Arranging for the sharing of that information

In practice, this involves separate technology and standards issues:

- Creation and maintenance of a Record Locator Service
- Definition of standards for secure sharing of information

We treat these efforts as separate for two reasons: first, we know from the growth of large technical systems in heterogeneous environments (e.g. email, the Web) that when too many separate standards are bundled together in an all-or-nothing package, the expense and organizational difficulty of upgrading everything all at once becomes prohibitive. Instead, we imagine a set of related standards that can be implemented in various orders, and can be effective at various levels of completeness.

Second, there would be advantages to either of these efforts in a clinical setting, even if there were little progress on the other. Finding a patient's records would be useful, even if those records ended up being faxed as hand-written notes. Simplifying information exchange among parties who need to share records, even if they are using inefficient methods to locate those records, would likewise be useful.

Construction of the Record Locator Service

The Record Locator Service is new infrastructure, and will require ongoing institutional support. In practice, it will be a cluster of databases holding four types of records -- patient identifying information, healthcare provider information, a list of patient records held by those providers (though not the records themselves), and contact details and other services made available by the providers.

The Record Locator Service (RLS) can only be used by authorized parties and only over secure connections, to allow a query to come in. Once the RLS receives an authorized query, it will search for a patient, and return a list of entities it knows have information on that patient, telling the querying institution where that information is located and whom to contact in order to access it.

Some organization will have to take responsibility for the ownership and operation of the RLSs; they will also be responsible for guaranteeing service level agreements, and must ensure the security and safe handling of the records contained in the database. There are a number of organizational models for this, from setting up a new institution who owns and operates the RLS on behalf of client organizations to a 'first among equals' approach, where an existing institution takes on the running of the RLS, in return for support from partners. The design of the institutional structure for supporting an RLS will be a key part of designing any pilot project.

The RLS would be queried when Institution A had a patient whose existing records they needed from other labs, hospitals, or clinics. Institution A would offer authorization credentials over a secure network connection. They would then send a request for records about a particular patient, offering a set of identifiers that uniquely identify that patient (e.g. name, DOB, gender, address, phone). These characteristics would then be run compared using the probability-weighted matching algorithm described above, with the locations of the matching records returned to the querying institution.

Recommendations for the Record Locator Service

Our recommendations for the Record Locator Service are as follows:

- Survey existing technical practices
- Survey approaches to distributed synchronization of databases
- Survey existing organizational arrangements
- Adopt or develop standard legal templates
- Launch pilot projects involving three or more entities

Survey of existing technical practices

During our work, we interviewed staff at several organizations and consortia that have patient populations numbering in the millions. This early work has convinced us that there is a wealth of valuable current practice in linking records across multiple data sources.

A second, deeper survey of existing practice is a vital next step, as we believe it will provide guidance to the construction of any future Record Locator Service by helping us identify practices already so widely adopted as to be de facto standards, as well as helping us understand what the significant challenges are likely to be.

Survey approaches to distributed synchronization of databases

Further study is needed on various parties' strategies for implementing synchronization and reconciliation of distributed databases. Because the Record Locator Service will need to be continuously available, all updates from participating organizations will necessarily be incremental, creating the risk of a mismatch between data held by Clinic A and the data held by the Record Locator Service on behalf of Clinic A.

Because there will never be a time when all data is deleted and reloaded, constant checking between local and remote records will need to be implemented in a way that maintains high data quality without creating unsupportable system load. There are several possible approaches to this problem; we will need to identify which of those approaches have been found workable by existing organizations.

Survey existing organizational arrangements

Any system that moves records across institutional boundaries involves significant organizational complexity. In addition to surveying current technical practice, a survey of current organizational arrangements is also vital. Our early conversations with the operators of large multi-stakeholder systems consistently elicited the same response: that arriving at the mutual agreement and contractual obligations among the participants was far harder than working out the technical details.

In order to get multiple participants in a healthcare network, there will need to be agreement on contractual obligations, articles of federation, funding arrangements, dispute resolution, and so on. Ideally, this survey would both inform and accelerate the design of standard contractual templates for use in future systems.

Launch pilot projects involving three or more entities

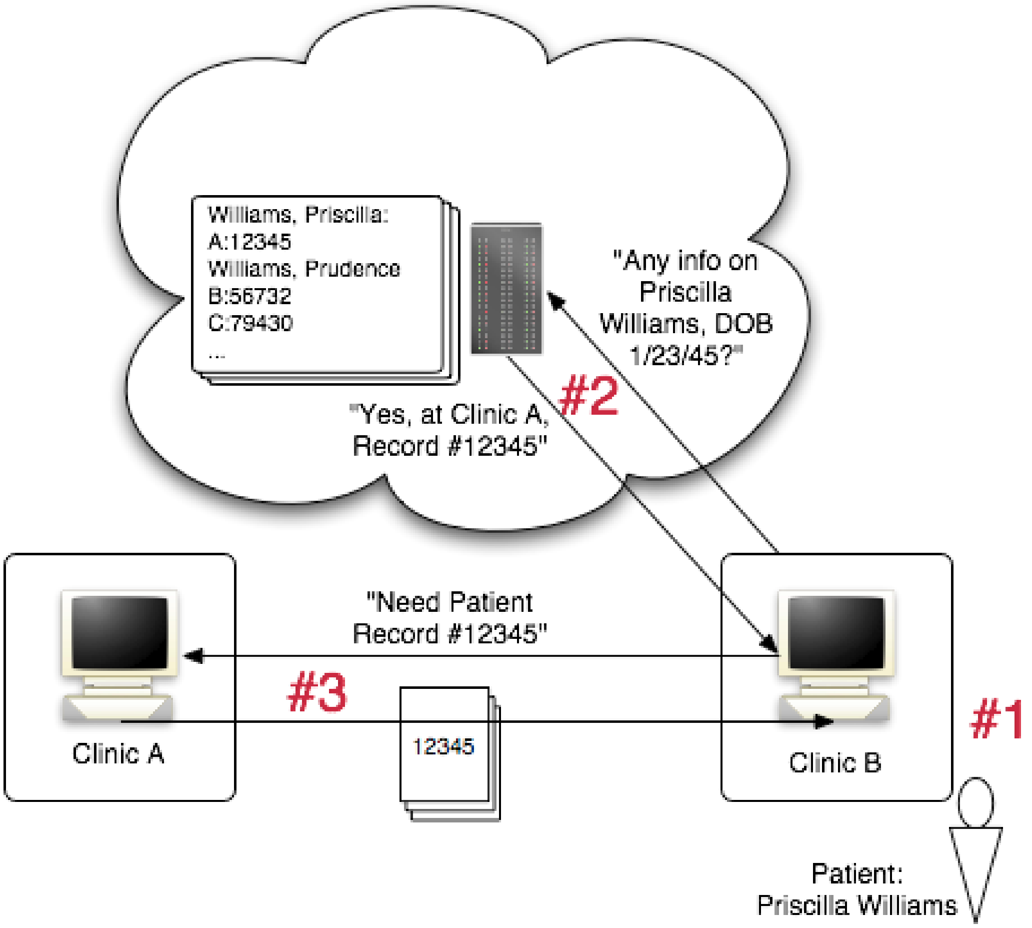
Because so many of the difficulties in getting any such system running are in negotiating multi-lateral agreements among the various parties, any pilot project designed to test the viability of the Record Locator Service must be multi-lateral, involving at least three parties at launch. Likewise, more than one of these pilots should be undertaken in the same time frame, in order to observe both similarities and differences between instantiations.

Sharing Appropriate Records

Assuming an Institution A queries the Record Locator Service, and Institution B holds information on the patient, A would then query B directly, at whatever level of technical sophistication is mutually possible. In an ideal scenario, both A and B would be able to share records automatically, with the records themselves also being expressed in a standard format. Should either A or B be less technically savvy than that (the norm today), they can default to less sophisticated tools, including the standard phone/fax round-trip.

Thus linking creates value in locating records, without requiring the immediate upgrade of every bit of health IT in the country. Likewise, upgrades (so long as they are interoperable) become more valuable as more entities begin exchanging records in this manner. This exemplifies our strategy in general: define a floor for technological engagement that maximizes participation, but provide every opportunity and incentive to outperform that minimum. Because upgrades in a system as large and fragmented as U.S healthcare will necessarily be piecemeal, the architecture needs to be a platform that both supports and rewards incremental improvements.

Example: Priscilla Switches Doctors


Williams, Priscilla:
A:12345
Williams, Prudence
B:56732
C:79430
...
"Any info on Priscilla Williams, DOB 1/23/45?"
#2
"Yes, at Clinic A, Record #12345"
"Need Patient Record #12345"
#3
12345
Clinic A
Clinic B
#1
Patient: Priscilla Williams

An illustration of how such linking, identification and sharing of a patient's records might happen: A patient, Priscilla Williams, moves and wants her new primary care physician at Clinic B to have the results of her most recent pap smear, currently held at Clinic A. If her new physician can't get the results, she will have to take the test again, resulting in additional expense, difficulty, and delay.

Clinic A, a participant in the system, has provided the Record Locator Service with an updated list of patients it holds records on. This is a background process, where Clinic A communicates directly with the Record Locator Service at regular intervals, rather than part of the individual search transaction.

Once the staff of Clinic B has taken Priscilla's identifying details (transaction #1 above), they will authenticate themselves to the Record Locator Service (RLS) to allow for auditing. After they are authenticated, they will make a request for the location of any of Priscilla's other records.

The request from Clinic B to the RLS will travel over secure transport such as a Secure Socket Layers (SSL). On receiving it, the RLS will compare Pricilla's information with its database. There are three possible outcomes here -- the Record Locator Service finds records with such a high probability match that they can be identified as Priscilla's; it finds no records that match; or it finds records that might match, and asks Clinic B for more identifying information. (This third option would require staff allocated to handling such requests; some system designs may simply treat such ambiguous pairs as non-matches, to minimize human input, even at the expense of additional false negatives.)

Assuming there is a match, the RLS will return pointers to other entities such as Clinic A that hold her records (transaction #2 above). Clinic B will then make a request for Priscilla's records directly to Clinic A, also via a secure Internet connection, again providing authorization credentials to show that it is allowed to do so (transaction #3).

Some of the resulting records may be returned from A to B directly over the Internet, using standardized interfaces for secure transport. The content of the messages may also be represented in a standardized format, for direct and automatic import into the new clinic's database, while other records may be sent by secure email, or even simple fax. Once B has the results of her earlier pap smear (as well as any other records held by clinic A), the staff of Clinic B can then add them to Priscilla's file.

Architectural Features

This example illustrates several key aspects of the imagined architecture:

- The main focus for use of the system is still in the hands of patients and providers – the system exists to support treatment, payment, and operations of the current healthcare system, rather than attempting to replace them.
- Even in its earliest form, it creates value for both doctors and patients. The staff of Clinic B can spend less time while gathering more information, Priscilla's doctor will be better informed, and both the doctor and Priscilla can avoid the expense and hassle of re-running tests that have already been done at Clinic A.
- It provides several layers of security. Only entities with authorization will be allowed access to the system; traffic between entities and network hosted services will be encrypted; no central repository of all identifiable clinical information will be held in the center of the network; and traffic between two entities will either be encrypted or take place outside the network (e.g. through fax or the mail).

- It does not require that Priscilla have any sort of national Health ID. Instead, it uses her existing identifying details to determine a match.
- It separates 'knowing that' and 'knowing what' information about the patient. The pointer database offers only 'knowing that' information, where information that a patient has records in a particular institution is available, but the records themselves are not.
- It leaves the information in the hands of the entities that have a direct relationship with the patients. Actual sharing of information is left between the requesting and responding entities, as today. The network service lowers the enormous costs and difficulties of discovery and location of remote patient records, but does not require the entities with the patient relationship to surrender control of those records to a third party.
- It leaves privacy controls where the information was created. If there is information Priscilla does not want disclosed, the institution holding that information can opt out of identifying Priscilla as a patient.
- It allows for enormous variability in the technical sophistication of participants. The minimum level needed to participate is a list of patients about whom an institution can provide records when asked by an authorized party. The providing institution at this minimal level of participation only needs to provide such a list, and to be ready to reply by fax or mail to valid requests, with no onsite technical requirements for hardware or software. At the other extreme, large multi-institution organizations can offer direct lookup and information sharing in response to authorized requests, thus potentially automating a complex and expensive task.

The above example illustrates a general design goal – create the minimal level of new functionality to be useful, while offering a gradient of services and automation that allows large entities with significant IT investments to gain additional value.

The basic threshold here is the ability to provide, in electronic format, a standard list of patients about whom a practice has information. This is the participation threshold -- "You must be at least this high to get on this ride." Institutions or providers that can't provide a simple list in electronic format are not ready to be members of the network. For others, the goal of connecting to the Record Locator Service may provide the impetus to clean up their internal record keeping.

Privacy Enhancing Technology Built into the Architecture

Any system of linkage or identification must be secure, preventing unauthorized outside access and limiting disclosures from within. Security and privacy policies and procedures that support electronic health information systems should provide strong controls throughout the environment and for all pathways and modes of access and use. Strong controls include a regularly updated authentication and authorization regimen; auditable records of access and transmittal; and mechanisms for enforcement, including sanctions for violation. No information system, regardless of the safeguards built in, can be 100% secure, but appropriate levels of protection coupled with tough remedies and enforcement measures for breaches can strike a fair balance.

Authentication

A critical component of privacy protection will be authentication of users. While further research needs to be done on authenticating users in large, decentralized systems, at this point a user name and strong password, properly managed, offer sufficient security. Proper management means, among other things, that procedures for issuing and revoking credentials must be strictly enforced. Since persistent identifiers pose a security risk, passwords must be time limited, so there is automatic revocation and reauthorization. In one major system we studied, passwords are issued initially in face-to-face encounters and are good for 90 days, and are reissued online with a “secret question” to verify identity.

This in turn raises the question of who manages authentication. There are several models, ranging from peer-to-peer to governmental, but the one that seems most likely to succeed (and that is most consistent with emerging practices) involves a non-profit entity at the center of the national system and at the center of the regional or other sub-network system that compose the national system. The New England Health EDI Network (NEHEN) is an example of such an entity at a regional level. It is a contractually based membership organization, supported by fees. Each member has one vote in system governance. The network admits institutions, and the institutions authorize individuals (doctors, nurses, administrators, etc.) pursuant to guidelines set by the network.

Authentication (who you are) is not the same as authorization (what you can access). Some hospitals permit all users to access all records. A better approach is to establish levels of authorization, at least based on occupational category or function. Under such an approach, doctors, for example, might get access to all records, while pharmacists would get access to one subset and administrators to a different subset.

The issuance and revocation of authentication in large systems is a specialized part of the computer security universe. Experts from that field must be brought into the design of the connecting for health system and its component systems.

Audit trails

Another important component of privacy protection is immutable audit logs for each access to a record, identifying the person who accessed the record and the purpose. Immutable logs are tamper resistant and tamper evident trails of activity. Ideally, these logs require multiple parties to access their contents, and all alterations are treated as updates; no data is ever deleted, and all changes are signed. In such logging models substantial collusion would be required to actually falsify the audit log. Such logs improve accountability and oversight, and can be used to identify patterns of abuse. Audit information would be available directly to patients, as well as to other participants in the health system.

Encryption

Encryption should be used to protect medical records both in transit and in storage. Virtual Private Networks (VPNs) or Secure Socket Layers (SSL) offer protection to data in transit. In addition to encryption of the data as it passes between entities, we considered the possibility of encrypting (or "hashing") the identifiers in the RLS. We believe that such hashing is a promising technique, especially when used to compare sensitive numerical data such as SSNs, but that it should only be contemplated when it can achieve substantially similar results as comparing unhashed data. We concluded that the greatest risk is not theft of database, but insider abuse by authorized internal users, so we focused protections on that problem.

Limiting queries

Further protections can be built in by limiting query formats. For example, the directory can be designed to make it impossible to ask for all 20-25 year old females in a certain neighborhood. In this regard, we believe it is important that Social Security Numbers (SSNs) not be used as search terms. The SSN has become so compromised as a result of its widespread use as a generic identifier that it is a risk to any system that relies on it. We recommend that health records systems wean themselves from the SSN as the patient record number, and that the SSN be reserved for disambiguation of records, possibly using only the last 4 digits, and that where possible, comparisons of SSN matches be conducted with hashed data, so that no SSNs are actually stored in the system itself.

Network of Networks

It's important to note that the proposed architecture sets a floor for interaction between entities needing to share clinical information, but not a ceiling. Much higher levels of interoperability are possible and, if agreed to by all parties, desirable.

As an example, there are a number of health initiatives that tie together several institutions into one network, including many of the organizations we surveyed such as CareGroup, NCEDD, Regenstrief, and Sutter Health. These regional networks have higher degrees of both contractual and technological standardization than specified here, and consequently offer a higher level of service.

Because of the significant value of these systems, we specifically imagine the architecture proposed here working as a network of networks. Not only can individual physicians and small or large institutions connect, but regional networks can connect as well. By adopting this network of networks model, we will be able to take advantage of the value of regional systems without having to re-build what they have done, and we will allow other entities to locate records held in regional networks without forcing them to admit new entrants nationwide.

Though we have benefited enormously in this work from examining these regional models, it is worth noting that they cannot simply be copied at larger scale to create a potentially national architecture, for several reasons:

- They are much smaller, involving a few dozen entities. This enormously reduces the complexity of the required infrastructure.
- There is a much higher degree of both trust and familiarity between entities in a regional network, which are likely to share care for many patients, and to refer patients to one another frequently. By contrast, any national system must work even between entities that don't collaborate or share information regularly.
- Regional networks typically have a high degree of mutual contractual obligation, including shared financial obligations. This is beyond anything that can be imagined at a national level -- provision must be made for simpler and less onerous obligations for participants.
- Regional networks typically operate in the borders of a single state, eliminating the cross-border complexities of varying state regulation.

By contrast, the system we describe here needs to work at a scale of many hundreds and later many thousands of physicians and institutions; among parties who rarely or never interact; across state borders; and without imagining creating a national health system by legal fiat.

We believe the network of networks model allows us to get the best of both worlds. As most care is local to the patient's home, they can receive care within the regional network, but when they need records moved outside that network, currently very difficult, this system will provide the necessary services, as well as serving as a method of connecting healthcare entities not affiliated with any regional networks.

Incremental Participation in Information Sharing

Given our belief in incremental development, proceeding on several fronts in a decoupled fashion, a key problem is going to be what technologists often call an "impedance mismatch," an analogy to the difficulty of connecting circuits with different resistance to current. This mismatch occurs whenever two systems that need to interoperate have different requirements or levels of technical sophistication.

One example of such a mismatch would be between entities that can and can't provide round-the-clock access to their records. Likewise, some institutions may expect near real-time communication with partners (analogous to instant messaging), while others may only be able to support asynchronous communication (analogous to email). In both of these cases, it may be necessary to provide an intermediate service to solve these mismatches.

Since the prime virtue of incremental development is to allow individual entities to implement or upgrade at their own pace, any system aiming for broad participation will inevitably confront the problem of entities interested in participating, but whose current IT systems are inadequate to the task. Without a solution, the difference between IT haves and have-nots will be exacerbated, not lessened.

Since the system will necessarily grow in pieces, particularly in the pilot program phase, many of the decisions about how to handle impedance mismatch will be best handled on a case-by-case basis. However, the Linking Working Group did identify two strategies we believe will be worth testing in these situations: local gateways, and proxy servers.

With the Record Locator Service in the center of the network, and the record-holding entities at the edges, these two strategies are listed in ascending order of centrality, which is to say distance from the edge entities and proximity to the Record Locator Service.

Local gateway

One of the simplest ways to provide for interoperability between a local institution and the rest of the network is to provide a low-cost gateway, a simple computer that sits between an institution's local network and the broader Internet, and is connected to both. The gateway would have three functions: First, it would provide a standards-compliant interface to whatever system the local institution happened to be running. In practice, this would mean fielding queries from the Record Locator Service in whatever format that service produces, translating those queries into queries for the local database, which might be as simple as an Access database running on a single PC. Likewise, it would take the results of such a query, wrap them up in the format the Record Locator Service expected in return, and send them back upstream.

This idea has a long pedigree, being the pattern the Internet launched with. In that situation, the original designers reasoned that it would be too hard to engineer interoperability between all computers at all the sites selected for participation, and instead provided small computers that acted as gateways between local systems and the Internet. This model is currently in use in the Regenstrief system in Indiana, to good effect.

Proxy servers

A second pattern is to provide proxy servers. Like the local gateway solution above, a proxy server also functions as a gateway between local and remote resources. Unlike the local gateway, however, it would not be hosted at the local clinic. Instead, it would be hosted on the Internet, where it could be made accessible to more than one institution. This would have the advantage of cost savings and ease of implementation, as several institutions or providers could be clients of one such proxy, sharing the costs. However, it raises the requirements for system-wide IT support, meaning such a solution would have to be useful for a number of participants, in order to spread the cost.

Recommendations for Information Sharing

In addition to the recommendations already made for further work on the Record Locator Service, additional work will need to go into arranging for subsequent transport of those records, including:

- Survey existing technical practices
- Adopt standards for secure transport
- Experiment with varying forms of proxying and caching

Survey existing technical practices

In addition to surveying existing practice for record linking, it will be important to survey existing technical practices for networked communication. These practices will include issues of identity, (how does an provider know who a request is coming from), authorization (how does a provider decide whether they have a right to see the records they are requesting), security (how can a provider respond in a way that protects patient information), and transport (how do providers communicating with one another -- over the Web, via secure email, etc.)

It will be essential to understand early on which of these areas have broad adoption or consensus, and which are still largely undefined, so that any future work can concentrate on the latter set of issues.

Define standards for secure transport

Whether or not there is consensus on secure transport for medical records, any future effort will have to specify acceptable communications tools (e.g. Web, file transfer protocol, email, etc) and minimum levels of security for those tools. Though we generally recommend keeping technical thresholds as low as practical, in order to maximize engagement with the largest possible proportion of healthcare institutions and providers, in this case adoption of and conformance to standards will be a necessary threshold to raise. Depending on the design of any given pilot project, it may be possible to share records securely by other means, e.g. fax, but under no circumstances should participants be allowed to use insecure Internet transport.

Experiment with varying forms of proxying and caching

Because the mismatch between various systems is such a pervasive problem, pilot projects designed to test sharing of records between institutions and providers should ideally include tests of local gateways or proxy servers, in order to understand the dynamics of participation between partners of varying degrees of technical sophistication.

System-Wide Concerns

Because the linking problem necessarily involves a network of participants, there is a set of concerns that are associated with the system as a whole:

- Interpretation of received records
- Defining membership
- Certifying Standards Conformance

Interpretation of Received Records

While an obvious goal of any attempt to improve healthcare through better use of IT will include wide support for electronic health records, our proposed principle of decoupled development suggests that participants should be able to share any health information they may have on file for a patient, in whatever format it exists. In practice, this will mean that at least some data is unstructured, possibly as scans of written notes, while at the other extreme, some data will be in a well-structured electronic health record format.

Thus, for any pilot project, there will be a range of support for the receiving party in interpreting or handling the health records themselves, ranging from no formal support to full computer-aided decision support:

No interpretive support

The most immediate benefit of being able to locate and share records comes when the information is communicated in any form (even fax) to a caregiver or other person that will interpret the information in order to make decisions.

Support for filing, retrieval and display only

An intermediate benefit occurs when the pair of communicating systems agree on a format that allows the information to be filed and retrieved in a manner that supports the user to see the information when needed, even if the body of the information is still in verbal or even "fax image" form.

Computer-assisted decision support

Still more benefits can be achieved when the pair of communicating systems have the internal capability to store data in finely structured, coded manner so that the computer can provide decision support.

> Decision support can be as simple as avoiding an order for a test that has already been performed to detecting drug allergies or to much more elaborate rules that support patient safety and evidence-based medicine.

Many systems already produce data in a manner consistent with computer-assisted support. These include laboratory systems and provider administrative systems and electronic health record systems for specific kinds of data. Likewise, the ability to create or use structured and coded data is not limited to the support for clinical systems. Many administrative systems produce or insist on receiving data in a coded or structured format.

The challenge is to create an environment that facilitates communications among pairs of systems at all three levels of sophistication and, indeed to facilitate communications among systems that operate at different levels. Our design is flexible in that it facilitates communication among end-point systems at varying levels of sophistication in the structured and coded representation of data. The design also supports the evolution of systems as they move towards increasingly well-structured data. For example, while some might use the environment to locate records and request them by telephone or fax, others may draw on it to support the full electronic exchange of highly structured data for sophisticated data analysis and decision support.

This is necessary because health information will continue to be a mix of unstructured and structured and coded data, so we will need to support all possible forms of sharing between end-point systems with varying levels of structured data. Ideally we should allow two end-point systems that support highly coded data to exchange it without loss of data, while a system that supports less or little coding should be able to receive information from systems with comparable levels of coding as well as from highly structured systems, and a system that supports a high level of coding should be able to receive, file, and make use of lightly coded data when this comes from another system.

The earliest progress can be made by assuming that all systems operate without formal interpretive support. There is a danger, however, that making that assumption becomes a self-fulfilling prophecy. Care must be taken in any system to build in incentives for continual upgrade of capabilities, especially those relating to the adoption of electronic health records, even when basic use of the system does not require it.

Network Membership

Though it would be possible to design a network in which all participants have a high degree of technical acumen, such a test would be a poor guide to the issues involved in a broader rollout. However, forcing any network to the level of its least sophisticated member will dampen much of the potential value.

One possible solution is a two-tiered system, with two classes of participants: members and users. The users will implement the absolute minimum amount of technology necessary to get records from the system (providing the authorization necessary to query the system while being audited), while members would implement the full range of technological requirements.

There will be several distinctions between users and members:

- Members will have a higher degree of contractual obligation to one another
- Members will have a higher degree of contractual operation to supporting the service, including timely provisioning of records, and of paying for at least part of its upkeep
- Members will agree to provide automated interfaces for secure sharing of records
- Members will agree to provide records in a standard EHR format
- Members will receive a higher level of service from the Record Locator Service and from one another

Users of the system, by contrast, will be authorized to query the system to locate patient information, but won't provide information themselves. The possibility of a lower level of interaction is included as an escape valve for entities that want to begin participating but which do not yet have the technical capabilities to offer member-level service interfaces. However, provisions may need to be made for use charges or other forms of offsetting revenue, to avoid having the economic free-rider problem turn the Record Locator Service into an unsustainable resource. In early pilots, these questions of levels of participation will have to be decided ad hoc.

Membership cannot be a one size fits all relationship -- there will be giant institutions, 30 bed hospitals, and solo practitioners participating, and the system must facilitate the sharing of information among them. (Indeed, coping with heterogeneous systems of varying levels of sophistication is one of the core challenges of any health IT system.) Instead, the contracting and fees will have to be negotiated to take into account the varying technological and financial circumstances of the members.

Every participant we've interviewed in multi-institution networks has emphasized two things relevant to membership. First, it takes significant time and effort to build the trust necessary to negotiate and commit to a multi-lateral agreement. Second, the uniformity of the agreement is itself a key virtue of the system -- it cannot simply grow as a series of bi-lateral agreements.

Certification for Standards Conformance

After some early test implementations there will be incentives to add software tools for certifying vendor products and their implementations. This will allow new participating systems to be added with less personnel time spent testing interfaces. The methods and software developed early on will have even more value as the system grows.

The next phase of this project should include consulting with the public and private sector organizations that have experience certifying compliance with standards. Some of the factors that must be considered in developing the software and methodology are:

> **Vendor testing.** Vendors should be able to use the certification tool from within their software labs to demonstrate that a given version of a product can be implemented in a manner that conforms with the profile. The certifying organization should award a certification identifying that a specific software version has passed testing for a specific use case.
>
> **Interface testing.** User organizations should be able to use the certification tool to demonstrate that a self-developed program or a vendor product has been implemented in a manner that conforms to a profile. The certifying organization should award a certification identifying that a specific identified system within a user organization has passed testing for a specific use case.
>
> **Automated testing.** The personnel of a vendor or user organization should be able to operate the test and receive certification without the direct involvement of personnel from the certifying organization except as necessary for technical support.
>
> **Remote testing.** The personnel of a vendor organization should be able to perform testing over the Internet without the need to travel to the certifying agency or have the certifying agency travel to the system under test.
>
> **Realistic testing.** When a software certification system is evaluating data sent by a system under test it is necessary to avoid the "minimal data loophole." This occurs because a transaction with almost no application data often appears acceptable. The certification scenario must evaluate incoming data to ensure that it represents the range of data that can be expected rather than minimally compliant data.
>
> **Inbound data testing.** When a software certification system is sending test data to a system under test, it is insufficient to determine that the system under test accepted the data without reporting an error. There must be an

evaluation that the data was faithfully entered into the database of the system under test. Performing this evaluation in an automated environment is a challenge that can be met to a limited extent.

Testing for error conditions. The software certification system must evaluate the system under test not only with correct data and operation. It must evaluate its response to incorrect data and to simulated errors in the network or communicating systems.

No performance testing. It is impractical for a certification service to verify the speed of a system under test without an elaborate and expensive test setup. This should be considered beyond the scope of the certification service.

Limits of certification. No certification service can guarantee that a certified system will comply with the profile for all possible cases. The testing process would require a prodigious amount of time by personnel of the vendor and user. Without people representing the certifying organization on-site there is always the potential that an unscrupulous person in a vendor organization could fake conformance by using special hacked software that is not actually released to production.

The benefit of good-faith participation in certification is that errors are caught at a time when they are easier and less expensive to correct. Good faith participation in certification benefits vendors and user organizations and should be regarded as a cost-saving and scalability measure.

Security

Security in any large network is a complex problem on its own. In particular, in the security domain, solutions to existing problems create new problems. In March of 2004, a security vulnerability on a type of firewall was discovered and, as is normal practice, the vulnerability was published so that it could be fixed by the owners of the affected machines. However, in less than 24 hours, virus writers had created a worm that exploited the weakness, releasing it before most of the target firewalls could be upgraded. The worm thus used existing security apparatus as a target for a successful attack.

This example, but one of many, illustrates the issue: security is a process, not a product. In a system whose contents are as critical as the imagined architecture's will be, and whose round-the-clock uptime is as critical as it will be, security must be both an early and steady concern.

Though many of the Linking Working Group members have or have had operational responsibility for secure systems, most of us are not security experts – rather, we have called on that expertise when building systems. That approach is needed here as well – a critical next step will be to convene a group of security experts to contribute to the architecture.

What follows is a brief outline of security features we know we will need; the work of the security experts in the next phase will be to both flesh out these intuitions and add what we have missed.

Our principal focus is on security services that support CAIN -- confidentiality, authentication, integrity, and non-repudiation.

> **Confidentiality**: Material existing within the system will only be disclosed to those authorized to have it, and who need it for treatment, payment, operations, or other authorized purposes.
>
> **Authentication**: The system will require presentation of identifying authorization for use, thus both deflecting unauthorized use and enabling auditing for monitoring of compliance with policy guidelines.
>
> **Integrity**: Material in the system will be defended against unauthorized alteration, and all authorized alterations will be logged.
>
> **Non-repudiation**: Transactions undertaken in the system will be acknowledged by both parties, and cannot be unilaterally revoked or altered.

Whatever the particulars of the security systems as they are deployed, they must serve those goals. Beyond that, the system needs security standards in three domains:

Wire security

Securing material “on the wire” means making sure that in its transit from point A to point B it is defended from eavesdropping, copying, or other interception. In practice, this means encrypting all the material passing over that connection.

As a result, we need to define a minimum set of security standards, so that all participants can know that their potential partners are also treating the material safely and a set of universally adopted security protocols, (e.g. the manner a virtual private network (VPN) is set up) so that even local experiments can later be connected into a larger whole without significant re-tooling.

In addition, there are some potential policy changes, as with the Medicare provision forbidding the use of the Internet to transmit information, that will need to be re-visited in light of a sound security policy.

Perimeter security

Securing material on the wire is only part of the answer – it’s no good securing material in transit from A to B, if B is the malefactor. As a result, we also need to secure the perimeter of the network as well.

In practice, this means requiring some form of authorization credential for anyone using the system for any reason, as well as an auditing program that allows use of the system to be evaluated later. (This function is analogous to the ‘black box’ airplanes carry.)

There is currently an enormous amount of work on secure federated authentication schemes, exactly what such a system requires. However, this work is in its early stages, and as any pilot programs are going to involve a relatively small number of participating entities, early tests will almost certainly use an access control list (ACL) strategy, where some authority issues usernames and passwords, and logs the system’s use.

This strategy is effective for small numbers of participants but becomes difficult as the system grows large. Thus we need to work on two tracks, securing near-term experiments and pilots with traditional access control solutions, while researching and experimenting with alternative forms of authentication in large, heterogeneous and federated environments of the sort any functioning architecture will exemplify.

In any case, whatever authentication method is used, the method will authenticate the participating organization or entity. This entity will, in turn, authenticate the individuals that it is accountable for, through employment or other relationships, who are to have access to certain sorts of records. A critical aspect of auditing is that responsibility is *not* transitive, but rests with the authenticated institution. An institution suffering from an exploit that allows unauthorized use of the system is responsible for the damage, even if the malefactor broke into the system, rather than being one of their authorized users.

Content security

Sometimes B is both authorized to use the system *and* a malefactor, as with the hypothetical case of a file clerk searching for his girlfriend's records, or the intern looking at the records of a famous patient. This type of attack can be limited by restricting what can be done with the information, even by authorized personnel, and by making sure that physical access to the equipment does not translate directly to access to its contents.

This translates into two tactics:

First, we have adopted what one of our members jokingly referred to as the "Karate Kid" maxim: "Best block is not to be there." The biggest risk in the system is to the servers holding aggregate data. A key security principle in this architecture is the de-coupling of identifying data about the patient from clinical information. Because the actual clinical records reside in the participating entities, the only material handled by the linking servers will be the fields needed to confirm a patient's identity, and to respond with pointers to his or her information residing on remote systems.

Second, and more speculatively, there is considerable work being done on on-disk encryption, where the contents of every file are stored in an encrypted format, and are only able to be decrypted when running in a trusted environment (e.g. with proper user authentication.) This means that should hardware containing information on patients be physically stolen, as happened recently in a system with more than a half-million records run by the records management company TriWest Healthcare Alliance, the material contained would not be available on re-boot without it being an inside job (which in turn significantly simplifies the task of identifying potential culprits).

Recommendations for Security

Because security is a process and not a product, it must be undertaken as an ongoing effort. Therefore, rather than producing several recommendations for security, we have a single one:

Form a security team to guide subsequent development

There is no way to fully secure a system in advance, and to secure it once and for all. All security involves a constant balancing of threat and the cost of deflection, the particulars of implementation, and the broader changes in the security landscape as new attacks and defenses appear. We recommend the hiring of a security team to provide ongoing advice to the creators of any reference implementation or pilot projects, and to test the results by attacking the system to test its resilience once it is operational.

Conclusion

We are optimistic that the recommendations provided here for improved linking of patient information can lead to marked improvement in the amount and quality of clinical information available at the site of care. However, it is always tempting to believe you know more than you do. In the planning phase of any technology project, this temptation shows up as a desire to predict in advance the results of proposed changes.

This is dangerous because such predictions are always in part wrong, and the larger the project and longer the imagined timeframe, the likelier it is that any error in prediction will be serious. The literature of large system design is filled with projects that wrongly assumed success would be an uncomplicated outcome of a set of proposed actions. In addition, our own conversations with groups successfully managing distributed multi-million record systems have convinced us that the myriad implementation details can only be dealt with in an operational environment or the closest possible simulation.

The best way around this dilemma is real-world experimentation, followed by updating the model in response, followed by more experiments. We believe the next phase of this work must make the design and launch of pilot projects a key goal.

The problems of something as large, distributed and heterogeneous as the US healthcare system are difficult to define, much less solve. In addition to the obvious difficulties of satisfying the needs of an enormous number of stakeholders, the interconnected nature of medical practice makes it difficult to cleanly separate problem. As a result, work on the linking problem quickly leads to a set of legal, architectural and organizational questions, and answering those leads to still further questions.

Even acknowledging these difficulties, however, we believe that significant progress can be made on the problem of linking patient information; that this progress can be made in a way which relies on existing technologies, is respectful of the patient's right to privacy, and does not require Big Bang development. Furthermore, we believe that the infrastructure that would be deployed to improve linking will provide considerable additional benefits by creating an environment in which use of electronic health information can flourish. We believe it is important to seize the opportunity to attack these problems, as an early step in the more general project of improving healthcare nationwide.

Appendix

MPI Survey Summary Report
Connecting For Health

August 2004

Survey conducted by Ben Reis, PhD, formerly with the Markle Foundation Health Program and Clay Shirky, Chair of this Working Group.

Overview

Healthcare organizations typically maintain a Master Patient Index (MPI) or Enterprise-wide Master Patient Index (EMPI), as the definitive listing of all of their patients. (We will refer to both classes of index as MPIs throughout.) All patient data stored by the organization is assigned a patient ID that can be looked up using the MPI. Two pieces of information concerning the same patient will (ideally) share the same patient ID, stored in the MPI.

In cases where information needs to be shared between multiple providers, or any other health organizations, it becomes necessary to match patient information across multiple MPIs. The same patient will usually be assigned different patient IDs in each organization's MPI. The patient's identifying information will need to be matched so that all the patient's information can be tied together.

Survey

The Connecting for Health Working Group on Accurately Linking Information for Healthcare Quality and Safety sought to understand the current practices and issues involved in locating patient data in systems with multiple entities, each with its own Master Patient Index.

To understand the issues faced by specific projects doing this today, the Working Group conducted telephone surveys with technical and administrative personnel at the following seven regional efforts:

- CareGroup, Boston, MA
- North Carolina Emergency Department Database (NCEDD), NC
- Provider Access to Immunization Registry Securely (PAiRS), NC
- Regenstrief Institute, Indianapolis, IN
- RxHub, multi-state
- Santa Barbara County Care Data Exchange, Santa Barbara, CA
- Sutter Healthcare, CA

These seven projects represent only a sampling of current ongoing efforts, and the present survey is an informational exercise, not a definitive scientific analysis.

Respondents were asked a series of questions covering technical, architectural, organizational and strategic issues surrounding the design, implementation and operation of their regional MPI based systems. A copy of the survey questions is available below.

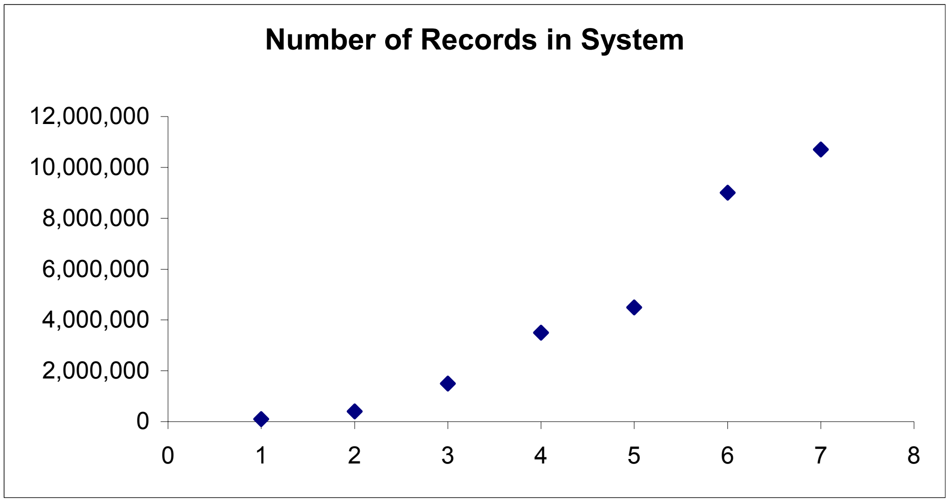
Overview of Results

Different projects currently being developed around the country are aimed at fulfilling a wide variety of different purposes. The designs of the various systems often reflect both the specific mission of the particular project as well as the organizational and technological conditions under which it was developed.

An overview of survey results is presented here. While the responses varied, some trends held true for most respondents.

Number of Records

The illustration below shows the number of records in each system. Most systems ranged in size from between 1 million and 10 million records. Santa Barbara had around 100,000 records. RxHub had over 150 million, being a combination of three largest Pharmacy Benefits Management (PBM) databases in the country.



The number of medical records in each system. RxHub, with over 150 million records, is not shown.

Number of Organizations

The number of organizations participating in each system varied widely. Usually, there are between 5 - 30 large participating provider sites (hospitals), but PAiRS has 400. Regenstrief includes an additional 1,000 participating small physicians practices. RxHub is much larger and had three large PBMs supplying data, with 10,000s of providers requesting data.

Organizational Structure

Respondents generally reported that their systems were organized in either a hierarchical or a peer-to-peer structure, with the organizational relationships defined by contractual and legal agreements between the parties. Some projects created an umbrella organization to both represent the parties as a collaborative and to operate the central MPI-lookup facility.

Technical Structure

Most systems reported keeping clinical data at the edges -- i.e. at the various local entities where it is collected -- with only demographic lookup for matching identity information across multiple MPIs being stored in the center.

Some projects maintain dynamically updating central copies of all the MPIs of the participating organizations. Queries are then performed across all the MPIs. Other projects merge these MPIs into one master MPI and queries are performed on the master MPI only.

Some organizations are actually moving towards a *more centralized* model, where all clinical data will be stored centrally. This offers major efficiency gains, as data synchronization between provider sites becomes much easier once all the data is stored centrally. Interestingly, in Santa Barbara, the central data hub is operated by a vendor, Quovadx.

Local MPIs

A vast majority of systems had one local MPI for each participating institution. Some systems had some local entities with no local MPI.

Data Quality

Most respondents indicated that data quality and cleanliness varied across different entities. Some indicated that hospital-based systems generally had higher data quality than those at smaller practices. Respondents found that they needed to focus on encouraging local entities to clean up their own data.

Most respondents indicated that small percentages of patients often do get assigned multiple IDs in their systems.

What Data is Shared?

Generally, respondents reported that all patient information present in the systems could be shared with other participants in the system, given that the appropriate privacy, security, authentication, encryption, etc. conditions have been met.

Discovery Process

While all systems automatically performed the multi-MPI identity matching step, the ensuing peer-to-peer data retrieval step was performed automatically in only some of the systems.

Other systems simply served to inform the user where the information was located, and it was the user's job to use other means to access it, including calling the particular institution by phone to request the records.

Fields Used for Identity Matching

When matching patient identities across multiple systems based on demographic information, most systems used some combination of the following data fields: Name, Gender, Date of Birth, Address, Phone, Zip, Social Security Number. Some systems applied different weights to different fields.

Dealing with Ambiguity

Respondents seemed to fall into three categories regarding the approach they take to handling cases of identity match ambiguity:

- Manual disambiguation at the center – Trained professionals working at the central site or at each institution use their own common sense and research to resolve ambiguous cases. This approach seems to be feasible for dealing with ambiguity in systems containing up to a few million records, where a few dozen specialists are able handle the load.

- Highly specific matching algorithm – This approach results in effectively zero false positives, declaring virtually all ambiguous cases as non-matching. The drawback of this approach is that many matches that would have been correct are left unmatched (many false negatives). This approach was especially beneficial to very large systems, e.g. RxHub, where manual disambiguation at the center is not feasible due to the scale of the system.

- Push ambiguity to the user – The user of the system is presented with the full list of possible matches, together with machine-generated probabilities for each match. The user then decides which matches are good and which are not. While this removes the need for central disambiguation, it requires everyday users to deal with issues that they would otherwise not have to.

Performance of Matching Algorithms

This varied widely, depending on which of the above approaches the system took. A number of systems reported using Initiate System's "Identity Hub" MPI matching product with positive results.

Security

Various combinations of the following security measures were reported: Authentication with username/password; authentication with physical token; encrypted storage; encrypted transmission; SSL/VPN; full audit trail; physical security of storage devices.

Caching Policies

Systems that do not store clinical data in a central location indicated that they do not cache clinical data there either. However, not all of them had a policy regarding the caching of data by the users requesting data from the system. Some do have a policy: any data retrieved by the system must be stored locally by the user who requested it, as if the requesting user had originally recorded the data him- or herself.

Anonymity and Pseudonimity

Most systems do not have a formal process to allow anonymous or pseudonymous records in their system. One respondent does have processes in place to allow VIPs to have pseudonymous records in the system that can be linked to the real record by authorized individuals.

Do Records Stay linked?

In some systems once records are linked, they stay linked. In other systems, they do not stay linked, and the linking must be re-done every time.

Biggest Development Hurdle

Projects reported significant challenges in the setup phase with data cleanliness and integration. Some also reported issues with the initial phase of legal efforts to work out the contracts among the various parties.

Biggest Operational Hurdle

Some projects are still dealing with data cleanliness issues. Others are facing challenges in growing their system to include more organizations. Others face political pressures with different stakeholders promoting progress in different directions.

Plans for Future Improvement

Plans for the future include getting cleaner data in more standardized form, from more organizations, and covering a wider geographical area. Plans also include providing access to patients through a patient portal.

Scaling

Most projects do not expect significant scaling issues. One project reported that it is expecting to scale up its capacity in order to be able to provide EMRs to local doctor's offices as an ASP service.

Standards Used

Many different standards are used, as appropriate for the data handled by the particular system. HL7 adoption is nearly universal. Some projects reported that they are looking to move towards a Web services model.

Contracts

Contracts are necessary to define the relationship between separate organizational entities. Since there are no standard contracts available for this relatively new type of cooperation, each project found that it needed to create its own.

Survey Text

As part of the Connecting for Health Working Group on Accurately Linking Information for Healthcare Quality and Safety, we are working to understand current practices and issues involved in locating patient data in systems with multiple organizations, each with its own Master Patient Index.

We'd be grateful if you could help us understand the work you've done on your system, in return for which we'll be happy to share the resulting Final Report with you when it's done.

Below are a few background questions – answers can be approximate. We will be calling to schedule a brief follow-up conversation.

General Notes:

1. How many organizations are in your system?
2. How many patient records are there in your (local) system?
3. How many in the system overall?

4. What does the organizational architecture look like? (Everyone operates as peers, everyone subscribes to some central organization, etc?)

5. What does the technical architecture look like? (Everyone operates as peers, everyone subscribes to some central database, etc?)

MPI

6. Does your (local) organization have a single MPI? Or multiple indices?
7. Do patients (intentionally or inadvertently) have more than one ID in your local database(s)?
8. Is the data from your partners about as clean as yours? More? Less?
9. How much data is shared? Full records? Just billing and administrative details?

Discovery

10. How do you discover that a patient has records in another system? (e.g. Query a central database? Broadcast a message to all other participants? Ask the patient to identify other places where they have records?)
11. What data do you use to determine a match? (e.g. Name, DoB, Gender? Social? Address, phone, email? etc.)
12. How well does this work? (e.g. Roughly what percentage of tested matches require further disambiguation?)

Security and Privacy

13. What steps do you take to safeguard patient data?
14. Are there any rules about holding or caching data from other entities?
15. Can your patients use the system anonymously or pseudonymously? How do you handle these cases (if you do)?

Subsequent Operations

16. Once two patient records are linked, do they stay linked, or do they need to be re-matched every time patient data needs to be retrieved across entities?
17. If they are re-linked, do the entities involved share indices? Or do they create a master foreign key? Held by whom?
18. Once the data attached to a patient is shared, is that data held in both locations, or deleted in the 'subscribing' institution to be re-imported later?

History and Future

19. What has been the biggest hurdle to overcome in getting to where you are to date?
20. What is the hardest part of operating the system today?
21. What is the biggest opportunity or priority for future improvement?
22. Do you expect to need to scale the system to a larger version than you have today? If not, why not? If so, what do you imagine the hardest coming challenge will be?

Post-script:

Obviously every system has a number of unique characteristics, most of which can't be captured in a few questions. In addition to the questions above, we would appreciate any information you can present on the standards and documents you use in your system. Our principal concerns are regarding technical, privacy and contractual standards.

A. What standards or format(s) do you use in expressing and sharing data (e.g. HL7, EDI, SOAP, etc.)
B. What standards or agreements do you present to patients vis-à-vis your HIPAA policies?
C. What documents or contracts set out rights and responsibilities among the various parties?

We thank you in advance for any insight you might be able to provide. We believe that better understanding real world cases will help us in making suggestions for a National Health Information Infrastructure that is compatible with current practice.

CONNECTING FOR HEALTH℠

MARKLE FOUNDATION *A Public-Private Collaborative*

Connecting for Health is an unprecedented collaborative of over 100 public and private stakeholders designed to address the barriers to electronic connectivity in healthcare. It is operated by the Markle Foundation and receives additional support from The Robert Wood Johnson Foundation. Connecting for Health is committed to accelerating actions on a national basis to tackle the technical, financial and policy challenges of bringing healthcare into the information age. Connecting for Health has demonstrated that blending together the knowledge and experience of the public and private sectors can provide a formula for progress, not paralysis. Early in its inception, Connecting for Health convened a remarkable group of government, industry and healthcare leaders that led the national debate on electronic clinical data standards. The group drove consensus on the adoption of an initial set of standards, developed case studies on privacy and security and helped define the electronic personal health record.

For more information, see www.connectingforhealth.org.

CONNECTING FOR HEALTH℠
MARKLE FOUNDATION A Public-Private Collaborative

Connecting for Health is an unprecedented collaborative of over 100 public and private stakeholders designed to address the barriers to electronic connectivity in healthcare. It is operated by the Markle Foundation and receives additional support from The Robert Wood Johnson Foundation. Connecting for Health is committed to accelerating actions on a national basis to tackle the technical, financial and policy challenges of bringing healthcare into the information age. Connecting for Health has demonstrated that blending together the knowledge and experience of the public and private sectors can provide a formula for progress, not paralysis. Early in its inception, Connecting for Health convened a remarkable group of government, industry and healthcare leaders that led the national debate on electronic clinical data standards. The group drove consensus on the adoption of an initial set of standards, developed case studies on privacy and security and helped define the electronic personal health record.

For more information, see www.connectingforhealth.org.