This product is part of the RAND Corporation monograph series. RAND monographs present major research findings that address the challenges facing the public and private sectors. All RAND monographs undergo rigorous peer review to ensure high standards for research quality and objectivity.

# IDENTITY CRISIS

## An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System

**Richard Hillestad, James H. Bigelow, Basit Chaudhry,**

**Paul Dreyer, Michael D. Greenberg, Robin C. Meili,**

**M. Susan Ridgely, Jeff Rothenberg, Roger Taylor**

Sponsored by Cerner Corporation, CPSI, Intel, IBM, Microsoft, MISYS, Oracle, and Siemens

RAND HEALTH

The research described in this report was conducted within RAND Health, a unit of the RAND Corporation, and sponsored by a consortium of health information technology companies: Cerner Corporation, CPSI, Intel, IBM, Microsoft, MISYS, Oracle, and Siemens.



*Cover design by Carol Earnest*



Published 2008 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

# Preface

A national health information network, or NHIN, that enables disparate health care information systems across the United States to allow authorized users to easily and quickly share critical health information has the potential to enhance safety and dramatically improve the quality and efficiency of the national health care system. A unique patient identifier (UPI) to use as a singular key to accurately link, file, and retrieve individual health records was seen as an important element of the national system and was mandated as part of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) legislation. However, privacy and security concerns about electronically sharing patient information have completely sidetracked the development of standards for a UPI and threaten to delay the development of the NHIN. This monograph examines the operational advantages and disadvantages, compares the errors, examines the costs, and discusses the privacy issues associated with the UPI and its alternatives.

This monograph should be of interest to health care IT professionals, other health care executives and researchers, and officials in the government responsible for health policy.

This research has been sponsored by grants from a generous consortium of health information technology companies: Cerner Corporation; CPSI; Intel; IBM; Microsoft; MISYS; Oracle; and Siemens. The right to publish results was retained by RAND. The research was conducted within RAND Health, a division of the RAND Corporation. A profile of RAND Health, abstracts of its publications, and ordering information can be found at www.rand.org/health.

# Contents

# Figures

# Tables

# Summary

Correctly linking patients to their health data is a vital step in quality health care. The two primary approaches to this linking are the unique patient identifier (UPI) and statistical matching based on multiple personal attributes, such as name, address, and Social Security number (SSN). Lacking a UPI, most of the U.S. health care system uses statistical matching methods. There are important health, efficiency, security, and safety reasons for moving the country away from the inherent uncertainties of statistical approaches and toward a UPI for health care. In this monograph, we compare the linking alternatives on the basis of errors, cost, privacy and information security, and political considerations. We also discuss operational efficiency, ease of implementation, and some implications for improved health care.

## Background

In 2004, the Bush Administration pushed forward the development of a national health information network (NHIN) to enable disparate health care information systems across the United States to be linked so that authorized users could share clinical information in real time. The many potential benefits of this network include significantly improved safety, quality, and efficiency of health services. In this effort, Washington joined many other governments that are pushing health care systems into the 21st century using Healthcare Information Technology (HIT). However, unlike almost all of the other governments, Washington is not developing a unique patient identifier to use as a singular key to accurately link, file, and retrieve individual health records.

Privacy and security concerns have completely sidetracked the development of a UPI for individuals in the United States, despite Congress's mandating in the Health Insurance Portability and Accountability Act of 1996 (HIPAA) that the Secretary of Health and Human Services (HHS) adopt standards providing for a "unique health identifier for each *individual*, employer, health plan and health care provider for use in the health care system [emphasis added]." Although an analysis completed for HHS in 1997 suggested a number of practicable options for a national patient identifier, subsequent hearings conducted by the National Committee on Vital and Health Statistics

(NCVHS, 1998) revealed significant concerns that the privacy and security of patient information could be threatened if it were networked beyond local health care information systems.

Very few comments at those hearings were directed specifically at the relative merits of the UPI as a patient identifier, but Congress subsequently prohibited HHS from expending funds in further study of a UPI without its explicit approval. This prohibition effectively stopped HHS from further considering or experimenting with a UPI as a means of linking health information in a national or regional network.

The effect of the congressional ban has meant that statistical matching schemes for identifying and accessing patient information are currently the only realistic option for developing the NHIN. But that is not to say that statistical matching, with its known errors and operational impediments, is, in fact, the best policy choice. The provider community is quite concerned about the errors and related patient-safety risks associated with relying on inherently uncertain statistical methods, particularly when the health care data are distributed across regional and national systems. These error-related risks are, consequently, an important barrier to an interoperable U.S. health care system and its potential benefits.

## The Errors in Statistical Matching

To locate health records, statistical matching attempts to string together enough identifying information about an individual to substitute for a unique personal identifier. It involves matching attributes, such as last name, first name, birth date, address or zip code, and gender, and it may use medical-record numbers and all or part of the Social Security number.

The problem with personal attribute keys such as name and address is that they are usually not unique to the individual, change over time, and are often entered into different systems in different formats. And data-entry errors, such as misspellings, add to the difficulties with this type of key. Repeated collection, distribution, storage, and use of these data also represent an important identity-theft risk.

Statistical matching can attempt to correct for some of these changes and errors: The most straightforward process is to tag all of the near matches for human resolution, or disambiguation. Such disambiguation imposes significant costs and operational inefficiencies, particularly if the physician must resolve the ambiguities. Advanced approaches "score" matches on "closeness" to the input set. Those with a high score may be accepted as a match. However, all such efforts are subject to the probabilistic errors inherent in statistical matching systems.

There are two types of errors—*false positives*, in which two different persons' records are declared to be a match, which can lead to such errors as the wrong patient's health data being obtained; and *false negatives*, in which two records for the same

person are thought to relate to different people, leading to such consequences as some of the patient's data being excluded. Both of these errors can lead to serious medical errors, waste (e.g., repeats of tests or the wrong tests), and considerable deviation from the promises of continuity and quality of care postulated for a connected digital health care system.

How frequent are these errors? Published data on duplicate or split records in medical-records databases (having some of a patient's health data stored under different personal attributes) indicate a high rate of such errors, averaging about 8 percent overall and trending higher for larger systems. This means that, 8 percent of the time, some health information about a patient will be missing when an exact match is required.

Although statistical matching attempts to reduce such errors, eliminating the errors will require the adoption of a UPI; without a UPI, medical records will continue to be keyed to changing personal attributes with frequent data-entry mistakes, perpetuating the error problem, especially in networking environments involving large populations and multiple sites of entry, as required for regional health information organizations (RHIOs) and the NHIN.

The most likely causes of false-positive errors are data-entry errors and use of an insufficient number of attributes in a statistical search for matches. Our analysis of an 80-million-record demographic database indicated that an error-free composite key made up of name, date of birth (DOB), zip code, and last 4 digits of the SSN would be required to unambiguously identify all patients. Removing the partial SSN from the key creates nearly 1,000 false-positive matches in this database. Larger health-record databases, such as those of a national or large regional network, almost certainly require a unique identifier to avoid false-positive errors.

In addition to the error rates that distinguish the two methods of patient identification, there are other, significant operational differences. The most important of these differences are disambiguation of uncertain matches, implementation of the matching alternatives, and the structure that is imposed on the NHIN by the methods.

*Disambiguation* is the process of resolving multiple potential matches into a match with the correct person. In general, statistical matching algorithms are likely to require substantially more-frequent disambiguation; often, disambiguation is done by human intervention. Many of the efficiency and safety benefits theoretically possible with HIT systems depend on eliminating such human involvement and its concomitant slowness, expense, and propensity for error.

It is likely that statistical matching algorithms will be required during any implementation of a unique patient identifier because of the need to match historical data with demographic information and because some people will not have a UPI available. An approach to this requirement is to add the UPI as another key for a matching algorithm. When the UPI is available, the link will be direct and immediate. When it is not, the matching errors will be no worse than those of the current process.

The complexity of identification using multiple personal attributes in a networked health care system increases dramatically with the number of personal attributes and the size of the network. Therefore, to decrease the complexity, matching algorithms probably require a hierarchical software architecture that manages identification. In this structure, regional Record Locator Services (RLSs) keep track of the patient records and patient attributes for each region. These RLSs would then interact with state and national RLSs to find the same person's records in another region. The need for frequent updates in the RLSs because of the nonpermanent nature of the personal attributes used in a match and the need for human disambiguation of close matches further constrains the architecture and efficiency associated with a matching algorithm in an NHIN. The UPI permits queries between providers directly and would not require the hierarchical RLSs.

## Privacy and Security with a UPI

Much of the controversy with the UPI has little to do with the identifier *per se* and much more to do with what the emerging NHIN architecture, connectivity, and interoperability would mean for privacy. Despite the halt in work on a UPI, HHS continues to develop the NHIN, arguably without adequately addressing the same privacy concerns that caused Congress to table the UPI in the first place.

A comprehensive analysis of UPI options, commissioned by HHS before the congressional ban, supported the use of a UPI and concluded that among its strengths was accurate identification without the "repetitive use and disclosure of an individual's personal identification information," thereby preserving anonymity, protecting privacy, and preventing unauthorized access to health information. While acknowledging the possibility of risks associated with misuse of the UPI, the author (Appavu, 1997b) stated, "since access to healthcare information is possible even without the use of a UPI, the solution to this and other legitimate concerns does not lie in eliminating the use of a UPI," and suggested that the threat of "rigorously enforced" legal sanctions would limit the potential for abuse. And, in contrast to using personal information, the ability to issue a new UPI should facilitate reestablishing security after a breach of a patient's health information.

Giving an individual a choice of whether to acquire a UPI could reduce overall privacy concerns. Those worried about misuse of the UPI could simply opt out. In a voluntary system such as that proposed by Hieb (2006) and the American Society for Testing and Materials (ASTM, 2000), many of the potential HIT benefits of continuity of care and efficiency might be achieved if significant participation of those individuals seeing multiple providers could be obtained, either through physician encouragement or through employer or insurer incentives.

Under current federal and state privacy rules, the proposed NHIN is likely to generate legal problems, regardless of whether it employs a UPI or statistical methods. HIPAA did not anticipate the development of fully interoperable networks, and the HIPAA Privacy Rules do not cover the full panoply of organizations that will be involved in collecting, processing, and using health records in the NHIN. On a somewhat different note, since medical providers have a legal obligation to take reasonable steps to ensure that recipients of protected health information do not violate the Privacy Rules, it follows that providers may become risk-averse about participating in future extended networks in which the opportunities for *due diligence* (e.g., confirming the identity of recipients, the validity of medical-record queries, and/or the security of distant network-access points) are far more limited.

In contrasting the relative merits of the UPI and statistical methods, we point out that a unique patient identifier, once developed, would immediately become protected health information under federal and (applicable) state law. UPIs would be sensitive information and could be a target for illicit access; however, unlike the demographic components of a statistical matching algorithm (such as the SSN), the UPI would not link to financial records, which are the specific target of identity thieves.

## The Costs of a UPI System

Costs of a national identifier are a concern. It might seem that we could use the existing SSN as a UPI. Since most people already have an SSN, since the Social Security Administration (SSA) is available to manage that system, and since the SSN already serves as the health care identifier for about 20 percent of the population, it is a logical choice. Unfortunately, the SSN has been recorded incorrectly in health care systems a high percentage of the time, because it does not have check digits[1] or other means of verification.

HHS (1998) published a white paper describing additional shortcomings. Probably the most important argument against using the SSN is that its wide use for so many other purposes has led to its being frequently compromised as a secure identifier. In fact, some states have adopted legislation that specifically prohibits the use of the SSN as a health identifier (Erickson, 2004; Milbourn, 2002). The important implication is that either a new, more secure ("enhanced") SSN must be issued or an alternative UPI promulgated, with potentially large costs of implementation.

Two significant components to the cost of this option are (1) the cost of reissuing the enhanced SSNs and (2) the cost of modifying all the myriad of existing automation systems so that they can correctly process the new SSN. The SSA has estimated

---

[1] *Check digits* are numbers that are formed from arithmetic operations on the rest of the identifier (ID) digits and then appended to the ID. By repeating the operations on the ID digits and comparing the ID digits with the check digits, an automated process can determine whether the ID has been entered correctly.

the cost to issue an enhanced SSN for 277 million current SSN holders as $3.9 to $9.2 billion (Appavu, 1997a), depending on the security features built into the new card. Alternatively, the National Governors Association (2006) has estimated that issuing a "Real ID" (i.e., one ID per person and one person per ID), based on the Real ID Act for establishing an authenticated identifier through the issuance of state driver's licenses, would cost about $37 per ID. The total cost of issuance ($11.1 billion for 300 million individuals) is consistent with the upper end of the SSA estimate.

Guaranteeing that everyone in the United States has a health identifier that is unique and canonical would likely cost about the same as the proposed Real ID system and would require a national infrastructure for support. These costs, although much smaller than the estimated $80 billion per year (Hillestad et al., 2005) in potential benefits of a connected, interoperable health care system, are an important barrier to a UPI, short of such a rationale as Homeland Security dictating the development of a national identifier anyway.

A *voluntary UPI system* as proposed by the ASTM would have an estimated cost of $25 million for the first five years for the national organization issuing UPIs, not counting the cost to providers and RHIOs for performing the registration and administering their databases (Hieb, 2006).[2] We estimate the cost of registering all people in the country this way to be about $1.5 billion ($5 each for 300 million individuals, based on 5 minutes of health-care-provider office time at $1 per minute).[3]

This patient identifier (ID), guaranteed to be unique, should avoid obtaining the wrong patient's information and gradually reduce the problem of split medical records, as more and more records are filed under the unique ID. Because the approach is voluntary, it will require a parallel statistical matching method for those patients who opt out of the UPI, but such a parallel system is probably necessary to accommodate the transition to a new identifier and for those people unable to present their identifier but needing health services. One approach is to include the UPI as a key in statistical matching. When it is available, the match is immediate and certain; otherwise, it is no worse than the current system.

This monograph also briefly scans the policy and political environment for a national patient identifier. The key political consideration for developing national patient-identification policy today, as in the late 1990s, is widespread public concern for privacy and security of personal health information that is found in medical records. Today, however, privacy and consumer groups have a much greater appreciation of the value of Electronic Medical Records (EMRs) and networking than in the late 1990s, when the UPI was initially being debated; and there is greater understanding of the

---

[2] In a personal communication between Richard Hillestad and the author of the ASTM proposal, Dr. Barry Hieb, Hieb indicated a revised cost of $12 to $15 million for five years.

[3] RHIOs will need to do this registration for the NHIN for any ID systems, whether or not a UPI is involved.

technologies and policy options now available that could ensure reasonable privacy and security for patient information in a networked environment.

It is important to note that the vast majority of individuals and organizations addressing the patient-identifier issue are focused on privacy and security concerns, not on a patient-identification system or standard; very few organizations have addressed the distinctions between statistical matching and UPI. If the methods were debated publicly, it is likely that the added privacy and security risks associated with statistical matching would become an issue. But without this debate, statistical matching has had the advantage of not requiring new national policy and has, therefore, avoided being judged under the bright lights of public scrutiny.

## Conclusions

### Broad Adoption of a UPI Should Enhance the U.S. Health Care System

Our analysis of the costs, benefits, and other aspects of a UPI described in this monograph indicates that a health care system in which every patient has a unique, nondisclosing patient identifier is clearly desirable for reducing errors, simplifying interoperability, increasing efficiency, improving patient confidence, promoting NHIN architectural flexibility, and protecting patient privacy.

### A Hybrid System Utilizing Both Statistical Matching and a UPI Will Be Necessary for the Foreseeable Future

Such a system will be necessary when only some of the population has a UPI (as in a voluntary system), during the implementation of a UPI (which may take a number of years), and when a patient cannot provide his or her UPI and health services must be rendered. Depending only on statistical matching will perpetuate errors in health-records retrieval because of its reliance on nonpermanent and non-unique personal attributes. One possible approach is to begin phasing in a UPI as an additional attribute in statistical matching.

### Security and Privacy Could Be Strengthened with a UPI

In the context of a networked health information system, security and privacy have much more to do with how access is managed and records maintained than with a specific identifier approach. Password protection and encryption of a UPI are relatively easy, whereas encryption of personal keys used in matching algorithms decreases the power of the algorithms. Repeated disclosure of personal information and linking that information to health information, required in statistical matching in a network, probably carry a greater security risk of disclosure

of sensitive information than a UPI. Using demographic matching may also make it more difficult to recover from errors. Once a person's health information is known and associated with the patient's personal attributes, the option of giving the person a new identity with a new set of personal attributes does not generally exist.

### Costs of a UPI Are Significant but Probably Much Less Than the Value Associated with Error Reduction, Efficiency, and Interconnectivity of the Health Care System

Costs depend on the scope of uniqueness and how strong and centrally managed the registration and *authentication* (verifying that a person is who he/she claims to be) processes are. To put the costs in perspective, previous studies of the value of connected Electronic Health Record (EHR) systems estimated a potential efficiency savings of $77 billion per year at the 90-percent level of adoption, with additional safety and health values that could double these benefits (Girosi, Meili, and Scoville, 2005). A one-time cost of $1.5 to $11.1 billion for a UPI, to remove the systemic errors in health-records retrieval, is small by comparison.

### Implications for Public Policy

In this monograph, we further elaborate the importance of a unique patient identifier as an enabler of efficiency, quality, and privacy in a nationally connected health care system. HHS has not funded any development work on the UPI since the late 1990s. Consequently, none of the NHIN-development contracts funded by HHS has employed a UPI approach, which limits a key purpose of the consortium process: to experiment with and develop the best approaches to interconnectivity and interoperability.

Privacy and security appear to be inadequate under current law and must be enhanced as the health care system becomes digitized and interconnected.[4] However, prohibiting development of a UPI actually sidesteps the larger problem: the development of a NHIN without first establishing a legal environment that best protects privacy while also encouraging the advances that interoperability of EMR systems between providers would bring to health care quality and efficiency.

Although it is beyond the scope of this monograph to suggest specific policy actions the government might take to ensure the privacy, access, and security of health care information, it is within its scope to recommend that Congress remove the current and clearly counterproductive constraints on HHS with regard to the UPI. Instead, Congress should be encouraging HHS to make a full assessment of the privacy, security, and operational implications of all the alternatives for linking patients to their health records within the NHIN.

These issues should be the subject of open study and debate in the vitally important process of developing the best interoperable U.S. health care system and reducing

---

[4] For a discussion of the inadequacy of current privacy protections for a NHIN, see Greenberg and Ridgely (2008).

the errors and inefficiencies in that system. Continuing *de facto* endorsement of statistical matching as the only practicable approach to linking patients to their electronic health records will inhibit the effective development of the national health information network.

# Abbreviations

| | |
|---|---|
| ACES | Access Certificate for Electronic Services |
| ACLU | American Civil Liberties Union |
| ACP | American College of Physicians |
| AHA | American Hospital Association |
| AHIC | American Health Information Community |
| AHIP | American Health Insurance Plans |
| AMIA | American Medical Informatics Association |
| ASTM | American Society for Testing and Materials |
| BCBSA | Blue Cross Blue Shield Association |
| CAQH | Council for Affordable Quality Healthcare |
| CCHIT | Certification Commission for Health Information Technology |
| CCR | Continuity of Care Record |
| CDC | Centers for Disease Control and Prevention |
| CDX | clinical data exchange |
| CHCF | California HealthCare Foundation |
| DOB | date of birth |
| EDI | Electronic Data Interchange |
| EHR | Electronic Health Record |
| EMPI | Enterprise Master Patient Index |
| EMR | Electronic Medical Record |

| | |
|---|---|
| GAO | Government Accountability Office (after 2004); General Accounting Office until 2004 |
| HHS | Health and Human Services (U.S. Department of) |
| HIMSS | Healthcare Information Management and Systems Society |
| HIPAA | Health Insurance Portability and Accountability Act of 1996 |
| HIT | Healthcare Information Technology |
| HITSP | Health Information Technology Standards Panel |
| HL7 | Health Level Seven |
| ID | identification, identified, or identifier |
| IDN | integrated delivery network |
| IHE | Integrating the Healthcare Enterprise |
| IOM | Institute of Medicine |
| IT | information technology |
| LRP | Lake Research Partners |
| MPI | Master Patient Index |
| NCVHS | National Committee on Vital and Health Statistics |
| NHIN | national health information network |
| ONC | Office of the National Coordinator |
| ONCHIT | Office of the National Coordinator for Health Information Technology |
| PBM | pharmacy benefits manager |
| PHI | personal health information |
| PHR | Personal Health Record |
| RHIO | regional health information organization |
| RLS | Record Locator Service [or System] |
| SSA | Social Security Administration |
| SSDMF | Social Security Death Master File |
| SSN | Social Security number |

| | |
|---|---|
| UHID | Universal Healthcare Identifier |
| UPI | unique patient identifier |
| USB | Universal Series Bus (Flash drive) |

CHAPTER ONE

# Introduction

In 2004, President George W. Bush created the Office of the National Coordinator for Health Information Technology (ONCHIT; later shortened to ONC) to "provide counsel to the Secretary of Health and Human Services (HHS) and Departmental leadership for the development and nationwide implementation of an interoperable health information technology infrastructure" (HHS, Office of the National Coordinator, 2004). With this action, the Bush Administration pushed forward the development of a national health information network (NHIN) to enable disparate health care information systems across the United States to allow authorized users to share clinical information in real time. The many potential benefits of such a network include improved safety, quality, and efficiency of health services (Hillestad et al., 2005). In this effort, the United States joined many other governments that are pushing health care systems into the 21st century using Healthcare Information Technology (HIT). However, unlike almost all of the other governments, the United States is not developing a unique patient identifier (UPI) to use as a singular key to accurately link, file, and retrieve individual health records (Lau, Bruun-Rasmussen, and Bernstein, 2006).

Privacy and security concerns have completely sidetracked the development of a UPI for individuals in the United States, despite Congress's mandating in the Health Insurance Portability and Accountability Act of 1996 (HIPAA, 1996) that the Secretary of HHS adopt standards providing for a "unique health identifier for each *individual*, employer, health plan and health care provider for use in the health care system [emphasis added]." Although an analysis completed for HHS in 1997 suggested a number of practicable options for a national patient identifier (Appavu, 1997a), subsequent hearings conducted by the National Committee on Vital and Health Statistics (NCVHS) revealed significant concerns that the privacy and security of patient information could be threatened if such information were networked beyond local health care information systems (NCVHS, Subcommittee on Standards and Security, 1998).

Very few comments at these hearings were directed specifically at the relative merits of the UPI as a patient identifier, and even fewer comments suggested alternative identifier models. With the exception of those who oppose any form of health care identifier, the debate over UPI in the late 1990s largely focused on the very real

concerns that standards, business practices, and federal and state laws lacked adequate privacy and security safeguards. Congress subsequently prohibited HHS from expending funds to develop a UPI without explicit congressional approval. This prohibition effectively stopped HHS from further considering a UPI as a means of linking health information in an NHIN.

Many believe that the prohibition will not be overcome until security and privacy concerns are adequately addressed. However, given the important role that a UPI could play in *protecting* privacy, it is difficult to see how such concerns can be addressed, or alternatives compared, unless the UPI or the study of UPI-related options is included in federally funded networks and studies.

The effect of the congressional ban has meant that statistical matching[1] as the process for identifying and accessing patient information is currently the only realistic option for the developing NHIN. And advocacy groups, such as Connecting for Health (2005), have settled on statistical matching as the pragmatic policy solution. But that is not to say that statistical matching, with its known errors and operational impediments, is, in fact, the best policy choice. The provider community, for example, is quite concerned about the errors and related patient-safety risks associated with not having a UPI (AHA, 2006; Stubbs, 2005).

In short, in the United States, the statistical matching method is gradually becoming the *de facto* industry standard for patient-record identification—but without rigorous analysis of the relative merits of that method against the UPI, without any formal policy decision to do so, and without any formal standards to support its implementation.

To provide a more factual basis for the public debate about this important element of HIT functionality and connectivity, this monograph analyzes and compares the UPI and statistical matching with respect to errors, operational issues, cost, privacy, and political and legal considerations.

## Methods

Our study broadly examined the issues surrounding patient identification in health care. The study included a wide-ranging literature and statute review, interviews with health and information technology (IT) practitioners involved with patient identification and exchange of health information, and discussions with key national provider, insurer, consumer, and privacy organizations.

To study errors in the identification approaches, we conducted our own analysis, utilizing a large demographic database, the Social Security Death Master File (SSDMF) (SSA, no date). We also used published data about split medical records

---

[1] We describe the statistical matching process more completely in Chapter Two.

(which result largely from filing an individual's health information under different names, addresses, etc.), as well as publications reporting the performance of various matching algorithms.

We developed a qualitative model of an integrated health care system that we used to assess comparatively alternative identifier methods by performance, privacy, security, access, and value.

Finally, we estimated the cost of implementing and maintaining various identification approaches, using data about the number of entities and transactions in a national system and applying published estimates of the costs of various identification systems.

## Organization of the Monograph

Chapter Two describes the primary identification approaches, and Chapter Three examines the errors in the approaches. Chapter Four discusses other technical and operational issues with the alternatives. Chapter Five discusses the privacy and security implications of the alternatives. Chapter Six estimates and compares the costs of the alternatives. Chapter Seven describes the policy and political environment in which the study's recommendations must be received, and Chapter Eight presents our conclusions and recommendations.

CHAPTER TWO

# The Primary Approaches for Identifying Patients and Linking Their Health Records

Historically, patients visiting health care providers in the United States identify themselves in person at the point of care and authenticate that identity by way of something they are (e.g., a recognized patient, personal signature), something they have (e.g., a picture ID, insurance card, confirming family member), and/or something they know (e.g., their name, address, doctor's name, and appointment time). Providers typically assign unique record locators (often called *medical-record numbers*) to the records resulting from these visits. Such record locators vary widely, from simple patient and family names to modified Social Security or insurance numbers, to provider-generated alphanumeric codes. Properly identified patients can approve the sharing of these medical records with other providers and insurers by signing an authorization form, clearly identifying the provider of record, the individual or entity to receive the record, and the boundaries or limitations on the information to be shared.

These three processes—authenticating individuals, unambiguously linking individuals to their records, and authorizing controlled access to those records—used with paper-based systems continue to be core functions in the age of electronic records. But implementing these processes create new challenges in cyberspace. For example, in cyberspace, physical face-to-face methods of identifying and authenticating patients, providers, or others logging onto a network no longer applies; methods of electronic identification and authentication are required. Likewise, knowing a patient's name or medical-record number from a single provider is not sufficient to unambiguously access that patient's records from other providers or a regional health information organization (RHIO); each entity may be using different numbering schemes or name constructions. Further, names and other demographic information change over time; the larger the network, the more likely it is that more than one person will have the same name and other demographic data. Finally, while paper records can be lost, stolen, copied, or spied upon, they are not generally viewed as being as susceptible to inappropriate access, compromised data integrity, or widespread unauthorized distribution as electronic records. New security measures are needed. IT proponents assure us that these challenges can be overcome, but doing so demands new solutions.

This project focused mostly on one component of these new challenges: defining the best electronic patient-identifier system for the purpose of sharing personal health

information through electronic information exchange networks. The choice of how to do so has implications for the privacy and efficiency of the health care system and the quality of health care itself.

A 1997 analysis of patient-identifier options described three categories of patient identifiers: UPIs (e.g., patient-identification number, Social Security number), non-unique patient identifiers (e.g., medical-record number, possibly with a provider prefix), and alternatives to unique identifiers (e.g., Directory Service; Health Level 7 Master Patient Index [HL7 MPI] Mediation; or a Core Data Element–Based Identification) (Appavu, 1997a). In analyzing the use of these identifiers for the purpose of widespread (national) networking, the author found that

> there are two different approaches to addressing the nation-wide access. The first one involves an MPI look up with the use of a Unique Patient Identifier for a match. The second involves the search of an MPI with a given set of demographic information. This method may utilize a weighting algorithm to help the search. The probability of success increases with the use of increased number of demographic characteristics.[1]

For the purposes of this study, we define these two approaches as (1) the "UPI" and (2) "statistical matching." Our analysis centers largely on these two approaches, although we will, at times, discuss variations of each of these two and the advantages and disadvantages of other approaches.

## Statistical Matching

To locate health records, statistical matching attempts to string together enough information about an individual to substitute for a unique personal identifier, or key. It involves finding agreement of attributes, such as last name, first name, birth date, address or zip code, and gender, and it may use medical-record numbers and all or part of the Social Security number (SSN). As the database of records gets larger, more personal attributes must be used to keep the key unique. In some current and proposed record systems, employer and/or insurer names and member identity numbers are also important record-matching elements. A nearly unique and relatively stable attribute, such as an SSN, helps reduce ambiguity immensely in large databases, although some state statutes and some developing industry norms aim to restrict its use as an identifier. The algorithms vary from requiring an exact match on a specific set of attributes (often used in smaller provider organizations) to more-advanced probabilistic pattern matching.

---

[1] This assumes that the data elements for matching are entered accurately.

The problem with personal-attribute keys, such as name and address, is that they are usually not unique to the individual, they change over time, and they are often entered into different systems in different formats. And data-entry errors, such as misspellings and number transposition, add to the difficulties with this type of key. Particularly problematic are some types of non-English names, names with punctuation (O'Malley), names with prefixes ("van," "de"), names with common nicknames (William and Bill), and names for which it is difficult to distinguish the first and last names.

Statistical matching attempts to correct for some of the changes in time and errors: The most straightforward process is to tag all of the near matches for human resolution, or *disambiguation*. Such tagging has both cost implications and inherent operational inefficiencies. Disambiguation also may not be feasible if the human does not have direct access to the person to ask clarifying questions. The algorithm can also refer to a database of common nicknames, it can score possible matches on the editing required to change one spelling into another ("edit distance"), and it can use sophisticated pattern-matching techniques to estimate when a set of attributes is "close" to the input set.

Advanced algorithms preprocess the health-records database to determine the frequency of every attribute and score the match according to the discriminating ability of the specific attributes of that database. For example, a match of the name Smith typically would not score nearly as well as a match of a less-common name. The scores can be used with threshold values of acceptance and rejection, as well as with regions of possible matches that can be adjudicated by humans. The limits of acceptance and rejection are often tailored to the specific use of the matching and size of the organization. However, all such efforts are subject to the random errors inherent in statistical matching systems. Setting the acceptance and rejection limits higher or lower affects false positive, false negative, and indeterminant results. Minimizing one type of error comes at the cost of increasing other types of error. We discuss this trade-off shortly in Chapter Three.

## Unique Patient Identifier

The direct method for linking patients to their data is to create a unique, nonchanging alphanumeric key for each patient and associate that key with every health record. Finding the patient's records anywhere within the health care system is then a matter of verifying that the patient is the person owning the key (authentication) and asking each health care system or provider in the *domain* (country, state, region, or provider) whether it has information associated with that key.

The American Society for Testing and Materials (ASTM, 2000) *Standard Guide* lists desirable attributes of a UPI, including that it be

***Unique***—Only one person would be associated with each identifier, although the scope of uniqueness might be regional, national, or even international. The scope has important implications for errors, effectiveness, cost, and management, which we will discuss later. It is the uniqueness that permits the collection and aggregation of health information into a complete care record.

***Nondisclosing***—It should contain no personal information. It should not, for example, include address or name information, which could allow the association of a data inquiry with a specific patient or reveal other confidential information. The compendium of personal attributes used in statistical matching clearly violates this attribute.

***Invariable***—The identifier should not change in a person's lifetime (except to correct a problem, such as identity theft). This invariability avoids one of the most serious problems in statistical matching—changes in some of the personal attributes, such as name and address, making it difficult to find previous records.

***Canonical***—Each individual should have only one UPI. Guaranteeing such singularity requires an error-free, central registration process or, at least, that a central repository of IDs and authenticating information be available to each registration agency or organization. Multiple UPIs have actually been proposed as a means of giving a patient control of disclosure, but they can also lead to fragmentation of the individual's health care data.[2]

***Verifiable***—It is possible to determine the validity of the number—generally through the use of additional *check digits*—numbers that must match some mathematical combination of the ID's remaining digits without additional information. Verifiability helps to prevent input errors, which are another major source of duplicate records.

***Ubiquitous***—Every patient should have one. This is difficult to achieve, particularly if participation is voluntary, but the alternative is a hybrid system, in which some patient data cannot be found using a UPI.

The ASTM *Standard Guide* (2000) lists additional desirable attributes of a UPI, including that it be used only for accessing health information, thereby reducing both the incentive for identity theft and the proliferation of the identifier for uses other than

---

[2] Hieb (2006) is a proposal for a Voluntary Universal Health ID (VUHID) system that would provide unique numbers at the request of registered RHIOs. The RHIOs and associated providers would register individuals voluntarily in this system. The registration numbers could then be used as an identifier for entry into any provider within the RHIO. Furthermore, they could be used to request health information to be forwarded to another RHIO or care provider if the individual required care outside of the registering RHIO.

health care and opportunities for disclosure. By limiting its use to health care only, security controls and management can be vested in organizations driven only by the needs of health care, and the scope of potential fraud and misuse can be more tightly circumscribed.

An obvious question is whether the Social Security number could serve as a UPI. Key advantages are that most people already have an SSN and that the Social Security Administration (SSA) is available to manage it. Furthermore, the SSN already serves as the patient identifier for about 20 percent of the population.[3] But, the SSN, when used in health care systems, is recorded incorrectly a high percentage of the time because it does not have check digits or other means of verification.[4] A white paper published by HHS (1998) describes the shortcomings of using the SSN as an alternative. Among other things, an SSN is easy to counterfeit because the allowable formats are known; some people legitimately have more than one SSN; some individuals share the same SSN[5]; SSNs are not immediately available at birth; and some people are not eligible for an SSN. Perhaps the most important reason that an SSN may not serve well as a UPI is that it is so widely used for so many purposes—financial systems, taxpayer ID, student ID, etc.—that it is compromised as a secure, private means of health care identification, and using it in health care invites further identity theft.[6] Some states have adopted legislation that limits the use of the SSN in health care.[7]

There are proposals to issue a more secure SSN,[8] and the availability of the Social Security infrastructure makes doing so a reasonable alternative. The SSA estimates the

---

[3] It is used by the Military Health System, Medicare, and the Veterans Administration (plus an unknown percentage of private health care providers).

[4] Initiate Systems (2008) reports that, in an examination of six Master Patient Index (MPI) record systems, there were more than 150,000 duplicate records. Of these duplicates, the SSN was not present in one of them 61 percent of the time; when it was present, it was discrepant with that of the other record more than 50 percent of the time.

[5] The Medicare identifier for the beneficiary consists of the wage earner's SSN plus a suffix to identify the relationship of the beneficiary to the wage earner.

[6] In September 2005, in testimony before the New York State Assembly Committee on Consumer Affairs and Protection, Barbara Bovbjerg of the Government Accountability Office (GAO, 2007) reported that

> Private sector entities such as information resellers, credit reporting agencies, and health care organizations generally obtain SSNs from various public and private sources. Large information resellers have told us they obtain SSNs from various public records, such as records of bankruptcies, tax liens, civil judgments, criminal histories, deaths, real estate transactions, voter registrations, and professional licenses. To gather SSNs from these records, resellers told us that they send employees to courthouses or other repositories to obtain hard copies of public records, if not easily obtainable on the Internet or public record publications.

[7] Erikson (2004) reported implementation of a new law prohibiting the use of the SSN as health plan member identifiers in response to concerns about identity theft; Milbourn (2002) reported that health insurance companies are barred from using SSNs on ID cards and accounts.

[8] The SSA only estimated the cost of replacing the current Social Security card with something more secure. That more-secure "something" involves putting more data on the card—e.g., biometric information—and allow-

cost of issuing an enhanced SSN for 277 million current cardholders to be $3.9 to $9.2 billion, depending on the card security features, and the issuance process would take up to ten years (HHS, 1998). The primary objections to this enhanced SSN are its potential cost and that it is not restricted to health care. It is also quite possible that using any modification of the SSN (such as an additional check digit) would be prohibitively expensive because of the huge number of deployed computer systems that would need to be updated. The cost of issuing the card might be much less than the cost to find and remediate all of the computer-system software predicated and formatted on the current SSN, including the multitude of uses outside of health care.

However, combined with adequate authentication, issuing, and management procedures, a unique health identifier that has the ASTM characteristics should provide an unambiguous, stable, secure, and efficient link to a patient's records.

---

ing "employers to access the number holder's biometric information from a central, SSA-maintained database" (SSA, 1997). The SSA did not estimate the cost to employers of accessing this central database and using this additional information.

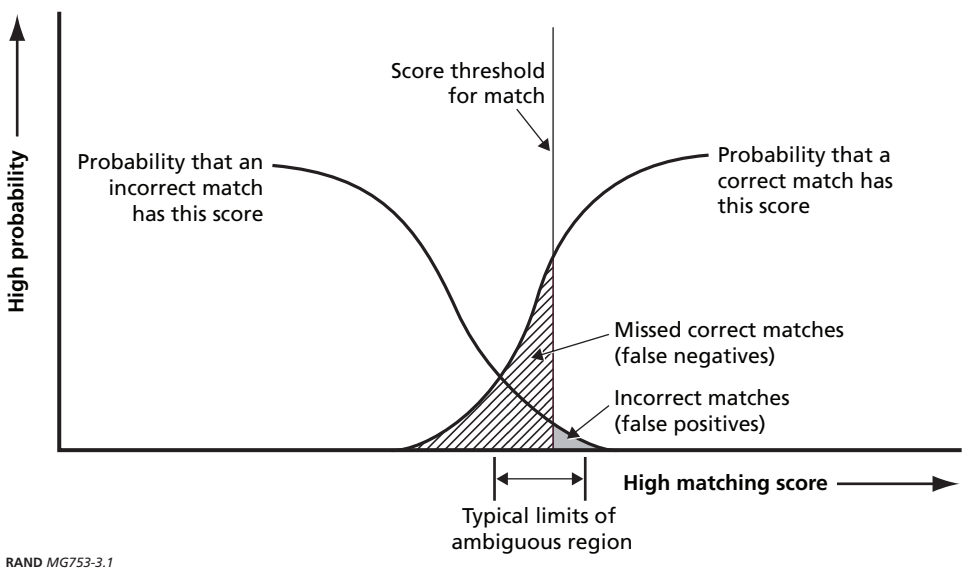
CHAPTER THREE

# Errors in Linking to Medical Records

The potential for error in statistical matching methods has important safety implications, which are a chief concern for many in the health care profession. Two types of errors are involved in statistical matching: *false positives*, in which there is a link to the wrong patient's records, and *false negatives*, in which not all of a patient's records are found. A graphic representation of these types of errors and of how they relate to the probabilities and threshold for matching can be found in Figure 3.1.

The horizontal scale shows the score of a particular match. As more and more attributes match and as the match is weighted by its score, or value, the higher is the probability that the patient is correctly matched to that record. A low score indicates a low probability of match (and a high probability that it does not match). It is possible

**Figure 3.1**
**Errors and Thresholds in Statistical Matching of Patients and Records**




RAND *MG753-3.1*

to use a threshold above which the record is assumed to match and below which it is not assumed to match, which leads to the shaded areas above and below the threshold. The area shaded to the right of the threshold is the region corresponding to false positives, or picking up the wrong patient's records. The shaded area to the left of the threshold is the region of false negatives, or the records of the patient that are not picked up because of some nonmatching personal attributes. Setting a balance between the two types of errors involves *tuning*.

Another approach illustrated in this figure is to define a region of ambiguity within which possible matches are tagged for human resolution, or disambiguation. Whether matching uses a single threshold or two thresholds, it is not possible to avoid encountering false-positive and false-negative matches. Adjusting the threshold or thresholds can result in a different proportion of false-positive and false-negative errors, but cannot be used to eliminate them because they result from the inherent characteristics of the population that lead to the two S-shaped curves.

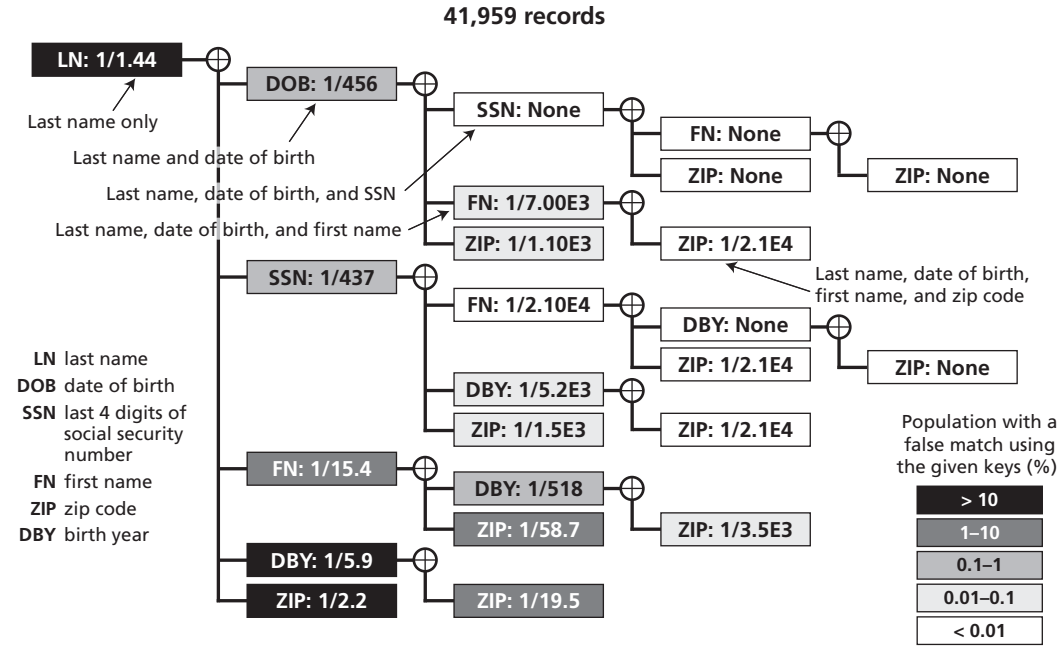
## False-Positive Errors—Linking to the Wrong Patient's Records

False-positive errors can be very serious. Obtaining the wrong health information about a patient can cause the wrong condition to be treated, the wrong operation to be performed, the wrong patient to be operated on, the blood type to be mistaken, erroneous lab results to be consulted, and the wrong medications to be issued. Is this, for example, the 40-year-old Mary Smith with diabetes and a penicillin allergy with her appendix intact or the 40-year-old Mary Smith in perfect health but with a removed appendix? Mixing these health records could have serious consequences. Sometimes these errors are obvious. However, when presented digitally, the data may be believed and acted upon, since the information may be intermingled with the patient's real data and indistinguishable from those data.

This kind of error is the result of accidental *record overlay* (more than one distinct individual assigned to the same record), health care ID theft, a threshold set too low, or a set of personal attributes used in the search that, in combination, are inadequately unique for the size and nature of the population being examined. Examination of medical-record databases by others has shown that the rate of accidental overlay is probably relatively small. For example, in two samples of about 5,000 records (relatively small) from two different hospitals, Grannis, Overhage, and McDonald (2004) found that the number of true overlays was 2 and 1, a rate of 0.04 percent and 0.02 percent, respectively.

An important cause of false positives is the use of an insufficient number of attributes in a search for matches. Figures 3.2 and 3.3 illustrate this problem. A large personal-attribute database of 80 million individuals, similar in size to a large RHIO or

**Figure 3.2**
**Chance of False Matches with Alternative Groups of Personal Attributes as Keys, for a Small Demographic Database**




SOURCE DATA: SSA, Social Security Death Master File (updated quarterly).
NOTE: The numbers in the blocks, such as the 1/1.44 in the leftmost dark block, mean that there is one chance in 1.44 tries of a false-positive match in this database when this type of key (LN, or last name) is used. Moving to the right in the diagram, the next block, "DOB: 1/456" means that there is one chance in 456 tries of a false-positive match when both last name and date of birth are used.
RAND *MG753-3.2*

state-sized records database, was used to create these diagrams,[1] the analysis for which is described in Appendix A.

In creating Figure 3.2, we used a 42,000-record subset of this database, similar to the size of a small hospital or large clinic.[2] At the top left, the result of matching only on the last name is indicated. For a random individual, there would be about a 2-in-3 chance (1/1.44) of finding another person's record with the same last name. However, if first name, birth year, and zip code are added, the number of possible false matches is reduced to only one in 3,500 (1/3.5E3). The use of a unique part

[1] We utilized a large demographic database, the Social Security Death Master File (SSDMF) (SSA, updated quarterly), to assess the probability of false-positive returns (finding the wrong patient) when various combinations of attributes, such as name, date of birth, zip code, and part of the Social Security number, are used as search keys and as a function of the size of the database.

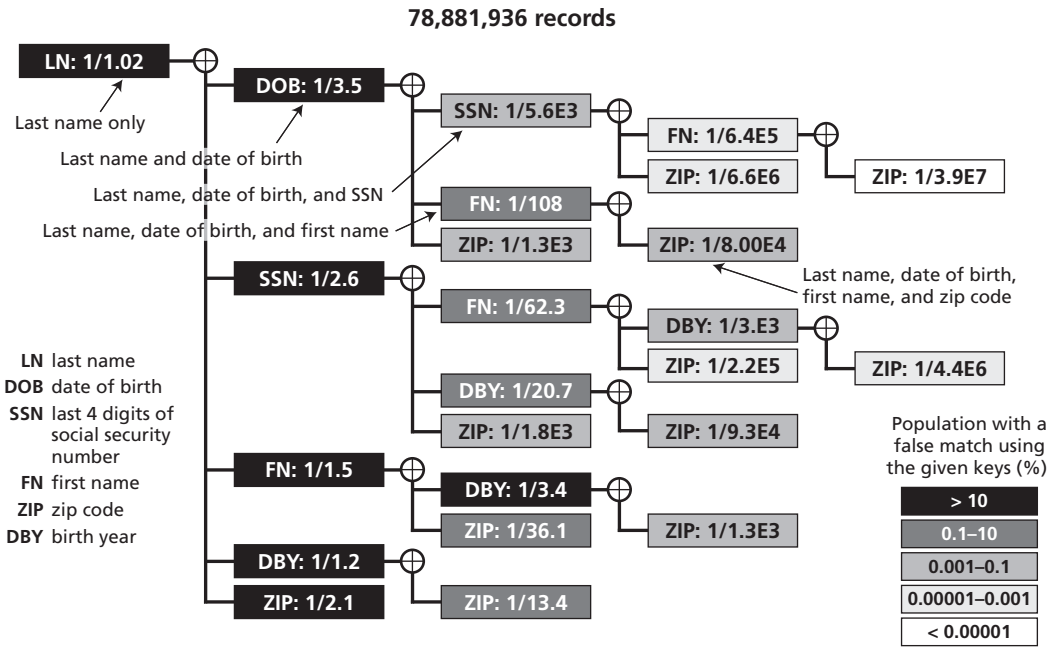
[2] The subset was created by selecting data from zip codes corresponding to a specific geographic region in Southern California.

of the SSN in the stream of keys quickly reduces the probability of a false match to near zero. This, of course, assumes that the keys for matching are entered correctly.

In the larger database of 80 million records, it is a bit more difficult to eliminate false positives. Referring to Figure 3.3, we see from the first block at the top left for a random individual for whom records are to be found that there would be a 98-percent chance that a false-positive match would occur with just the last name. When compared with the figure of roughly 66 percent for the small-population analysis, this shows how the false-positive rate is sensitive to factors such as population size. When date of birth is added to the key, the chance of a false match drops to 33 percent.

And, finally, after the last four digits of the SSN, the first name, and the zip code have been used to form the composite key, the rate of false positives drops to

**Figure 3.3**
**Chance of False Matches with Alternative Groups of Personal Attributes as Keys—Large Demographic Database**




SOURCE: SSA, Social Security Death Master File (updated quarterly).
NOTE: See Figure 3.2 for an explanation of how to read this diagram.
RAND *MG753-3.3*

1 in 39 million.[3] One conclusion is that, with enough correct personal-attribute keys, the false positives can be controlled to occur with very low probability.[4] However, eliminating the almost-unique SSN key dramatically increases the false-positive rate. If the database gets much larger, as in an NHIN, additional attributes or some, almost-unique, key, such as the SSN, is certainly required to keep this error rate small.[5] Thus, the uncertainties in statistical matching will always exist and the matching thresholds and related keys will need to be tuned for each Electronic Medical Record (EMR) database to keep such errors to a very low probability. If the use of an SSN as a key is ruled out, as it increasingly appears to be in many applications, ensuring a low rate of false-positive errors becomes quite difficult in such large databases.[6]

If the registration process is sufficient to guarantee uniqueness and the authentication process ensures that the UPI belongs to a certain patient, a unique patient ID by definition avoids this type of error. We note that it is necessary that the UPI be unique in the sense that this number is never issued to any other patient; however, to avoid a false-positive error, it does not have to be the only one owned by the patient. Thus the ASTM proposal for a UPI that permits a patient to obtain multiple IDs for privacy purposes would also help to eliminate such errors.

## False-Negative Errors—Not Finding Some of a Patient's Records

False negatives represent a fragmentation of a patient's health history, and not finding all key elements of a patient's record can lead to missing or incomplete information about medical conditions, previous surgeries, medications, or allergies, in turn leading

---

[3] To follow this path in Figure 3.3, choose the topmost boxes in the figure. Actually, this is a match of just two records in the entire database and is probably a result of an overlay in the SSDMF rather than a key that is not quite unique.

[4] Of course, the more data elements that are required increases the chance that one or more will be in error, which can adversely affect matching. A single erroneous search key could lead to the requirement for multiple additional accurate keys to avoid error. Errors in the search keys can lead to either false-positive or false-negative matches. We did not simulate errors in the search keys in our analysis.

[5] There is an analogy to this approach called the "Birthday Problem." Birthdays might be considered random keys that are assigned to every individual. It can be shown mathematically that, if you have more than 23 people in a room, there is at least a 50-percent chance of two people having the same month and day of birth. In this case, ignoring leap year, there are 365 distinct keys. If you ask the same question for birth minute, which increases the keys to 525,600 (there are more than a half million minutes in a year), it takes 854 people in the room to have a 50-percent chance of a match. Now, if there were 100 million people and you wanted to keep the probability of two persons' matching to 1 in 1 million, you would need a key whose size is 1,023, because the keys are possibly repeated among people. A unique identifier requires only 100 million (108) keys to achieve the same result.

[6] According to Connecting for Health (2005b), RxHub, a database of over 150 million records, reportedly achieves a very low rate of false positives with random matching, but it does so by "declaring virtually all ambiguous cases as non-matching. The drawback of the approach is that many matches that would have been correct are left unmatched (many false negatives)."

to possible life-threatening treatment errors and potential lawsuits. Missing information can also lead to inefficiencies, such as the cost of reordering of diagnostic tests and of delays and errors in treatment. Such inefficiencies have been estimated to cost the health care system more than $8 billion annually (Girosi, Meili, and Scoville, 2005). And analysis of longitudinal patient data for research or clinical quality and process improvement is much more difficult when some of the patient data are not found because of such fragmentation.

False negatives may be the result of changing personal attributes, such as name or address; of keying errors; and of changes in format, such as the order of first and last names. The order of first and last names is not obvious for some ethnic groups. Children are especially at risk for these types of errors, because prenatal tests and care are often recorded under the mother's name, a temporary birth name is often given at birth, and adoption or name changes are more likely to occur. All of these situations can cause the recording of some of the patient's data as new records, effectively fragmenting potentially important health information.

*Duplicate-record rates* are the frequency at which records are found that falsely appear to be those of another patient when in fact they should be identified as belonging to the reference patient. Published rates of errors in medical-record databases indicate a high rate of such duplication. One study of 112 Master Patient Indexes found that one-fourth of MPIs had duplicate-record rates of 10 percent or more, and only two had a duplication rate below 2 percent (Initiate Systems, Inc., 2008). The mean duplication rate was 8 percent. Larger files of 1 million or more records had a mean error rate of 9.4 percent, whereas those with less than 500,000 records had a mean error rate of 7.2 percent.

*Enterprise MPIs* (EMPIs), the combined MPIs of multiple organizations, had even higher duplication rates (7.3–39.1 percent in the 11 EMPIs examined) because of the additional chances of mismatched formats and inadequately merged data (Initiate Systems, Inc., 2008). It was found that female last name, SSN, and telephone number were the most common discrepancies among duplicated records. Corroboration of the magnitude of the errors is given by Tom Doyle of the Hospital Corporation of America, who is quoted as saying,

> Outside of carefully controlled pilots, accuracy for the current process (statistical matching) is roughly 90 percent, based on our collective experience and industry estimates . . . That margin of error will only widen as it is applied to ever-larger populations (Alliance, 2007).

It is important to note that such errors are not directly a result of inadequate matching algorithms. Indeed, the goal of the matching algorithms is to reduce these false negatives while not increasing the false-positive rate. Grannis, Overhage, and McDonald (2004), in an experiment with the two sample medical-records databases mentioned above, showed that some matching algorithms could reduce the false nega-

tive rates from 9–15 percent to 2–5 percent while increasing the false positives by only a small amount.

But the false-negative errors are related to patient-identification and record-matching methods. Eliminating such error completely will require the adoption of a UPI, whereas relying solely on statistical matching does nothing to resolve the data-quality issue and perpetuates the problem of record duplication/fragmentation, especially in networking environments involving large populations.

A well-implemented UPI would solve this problem for new records, but it would not resolve the fact that duplicate records already exist. For these duplicates, either a massive record cleanup and UPI insertion would be required or matching algorithms would need to coexist with the UPI. Furthermore, the UPI must be issued and used in a way that does not cause additional error, which implies that a strong authentication process involving multiple items of identification must be in place, as well as a means to detect and avoid errors in the UPI entry, such as entry by card swipe and check digits embedded in the UPI. Finally, to avoid duplicate records, individuals should have only one UPI—one potential problem with the ASTM proposal mentioned earlier. That proposal would permit the issue of multiple health care IDs to individuals, either because the individual registers at multiple RHIOs or for privacy protection. Multiple UPIs per patient, if truly unique, will not generally lead to ambiguous matches; rather, they will result in some information about the given patient not being retrieved if all the patient's UPIs are not presented.

Given the requirement for accurate authentication for a UPI, one might ask whether standardization of the personal attributes used by matching algorithms and similarly strong authentication would resolve the errors. Unfortunately, although doing so might help, the fact that the attributes change frequently over time and are not unique means that the errors would not disappear with such standardization. Adding the SSN as a matching data element would certainly reduce this error rate, but it also potentially creates more confidentiality and security issues than would the creation of a new UPI.

CHAPTER FOUR

# Operational Issues

In addition to the differences in error rates that distinguish the two methods of patient identification, there are other, significant, operational differences. The most important of these are *disambiguation* of uncertain matches; the difficulty of *implementation*; how much *architectural flexibility* they provide for structuring an NHIN; finding records for *patients without a UPI*; and *research and public health implications*.

## Disambiguation

Ambiguity in patient identification results when more than one set of records are a possible match for the patient in question. Ambiguity can occur when the identifier used to match is not sufficiently distinct to *uniquely* differentiate one person from all others represented in a database. In turn, ambiguity can lead to both false-positive and false-negative errors. As such, ambiguity in identification matching can be an important threat to patient safety and quality, as well as a source of inefficiency.

*Disambiguation* is a process through which multiple potential identification matches are further parsed until the patient can be matched with his or her data with sufficient certainty to allow for the delivery of a health service with reasonable confidence. The complexity of disambiguation varies according to factors such as the number of potential matches and the type of information available for further analyses. When sufficient digital data are not available to further differentiate potential matches, automated disambiguation may not be possible and may require human involvement.

Disambiguation entails implementing significant new workflows and may require substantial time and resources. When human involvement is required, many of the potential benefits of automation are lost. For example, at the point of care, disambiguation is often done by asking the patient further questions regarding personal characteristics and/or health care history. In some situation, disambiguation may not be possible, as when the patient is not present and information needed to further facilitate matching may not be accessible.

UPIs and statistical matching algorithms can differ substantially in their need for disambiguation. UPIs can be designed to be unique for each individual patient.

Statistical matching based on personal characteristics join data elements to create combinations that can differentiate individuals from one another. The performance of such matching is thus influenced by the demographic characteristics in a region and the number of possible matches. Therefore, identification approaches based on statistical matching will likely require significantly more disambiguation than UPI-based methods because of the interaction of the thresholds for avoiding false positives and false negatives as shown in Chapter Three.

Disambiguation, particularly disambiguation requiring human parsing, will have important repercussions for implementing Healthcare Information Technology (HIT) and the potential benefits HIT systems can provide. Many of the efficiency and safety benefits theoretically possible with HIT depend on eliminating human involvement (and its concomitant slowness, expense, and propensity for error) in data-exchange workflows. In addition, our studies have suggested that many of the benefits of HIT depend on the level of interoperability achieved, with much of the benefit depending on implementing machine-level interoperability (Hillestad et al., 2005).

If patients cannot be unambiguously identified via a computer-based process, machine-level interoperability will be hampered significantly. In turn, automating health services through the use of interoperable HIT systems may be limited or not possible at all. For example, if an outpatient laboratory attempts to send an emergency result to a physician's Electronic Health Record (EHR) in order to trigger an automated decision-support alarm, the laboratory system must first be able to uniquely identify the correct patient in the physician's EHR. If multiple potential records are returned that cannot be further disambiguated, then the laboratory would likely need to call or fax the physician's office, decreasing the potential benefits in safety and efficiency that the HIT systems could otherwise produce.

A central problem in the U.S. health care system is its fragmentation and high volume of transactions. High levels of interoperability would allow the opportunity for health care to be delivered through a very different paradigm, one in which health services are linked in chains of automated transactions—for example, entry of an electronic prescription leads to a mail-order pharmacy's checking eligibility and formulary data for the patient and mailing out the medications through a series of linked computer-to-computer transactions. Ambiguity in the patient's health care identity may prevent these digital health services from linking together properly, reducing the efficiency gains possible with HIT and increasing the risk of errors, such as delivering medication to the wrong patient.

## Implementation

HIT implementation is a complex process. On the one hand, statistical matching algorithms may decrease implementation complexity for identification systems and

for HIT networks. On the other hand, issuing a UPI for health care may require new aspects of health care infrastructure to be developed to handle patient registration, initial authentication, and UPI issuance. The quality of the UPI system will be highly dependent on the initial registration process, wherein the patient provides some type of existing identity credentials (e.g., birth certificate). If this step is flawed, all subsequent authentication steps may also be prone to errors. However, as demonstrated by the controversies surrounding the Real ID Act[1] (National Governors Association, 2006), setting up a system for verifying identity and then issuing the ID numbers would likely be complicated and subject to difficult political and financial constraints.

For the identification process to function, statistical matching does not require a new registration process or a new piece of identifying data to be issued. Moreover, statistical matching may better support the gradual migration toward an IT-rich health care system by allowing additional data items to be linked to a patient as they become available in digital formats. Data generated during ongoing episodes of care could still be retrieved via statistical matching approaches. By contrast, UPIs would require some type of organizational or governmental infrastructure to distribute the new identifiers and to manage the registration process before any data could be linked. During the migration period, some type of statistical approach may be necessary, particularly if HIT systems are adopted faster than the UPI infrastructure develops.

An important implementation issue is how historical data are handled. Unless historical data and records are linked to newly issued UPIs, UPIs would not be able to be used to retrieve existing digital data. Given the resources that would be required to label existing records with a UPI, such linking is unlikely to take place. Most likely, statistical matching would be needed to retrieve data from historical records.

Likewise, statistical matching algorithms may be needed to support patient choice and preferences. Participation in any UPI systems would likely require opting in or opting out of the system. For patients who decide not to participate in the UPI system, a secondary identification system would likely need to be developed and somehow integrated with the UPI system. This secondary system would likely involve patient demographic data and utilize statistical matching. *One approach is to add a UPI, when available, as an extra attribute key in statistical matching. When available, it provides a direct, exact match. When it is not, the matching errors are no worse than those of the current process.*

---

[1] The Real ID is basically a more-secure driver's license with biometric data (e.g., fingerprints or iris scans) and improved authentication for issuance. In addition to concerns about privacy, the governors objected to the costs to the states of implementing the Real ID.

## Architectural Flexibility

The complexity of identification using multiple personal attributes in a network increases exponentially with the number of personal attributes and the size of the network. Therefore, matching algorithms probably require a hierarchical NHIN software architecture that manages identification to decrease complexity (by preprocessing health-records data and creating regional and national Record Locator Services [RLSs]).[2] Currently, RHIOs are envisioned to be the hierarchical hubs in a potential NHIN, even though financial viability of existing RHIOs and, hence, of the architecture, is far from ensured (California HealthCare Foundation, 2007a). Also, the need for frequent updates in the RLSs because of the nonpermanent nature of the personal attributes used in matching and the need for human disambiguation of close matches further constrain the architecture and efficiency associated with a matching algorithm in an NHIN.

In contrast, a UPI permits queries between providers on a peer-to-peer basis and would not require the hierarchical Record Locator Service structure, allowing flexibility in the structure of an NHIN. The unchanging characteristics of a UPI also eliminate the requirement to know how and where data are stored so that record locators need not be updated as personal attributes change.

## Patients Who Do Not Have a UPI

It is unlikely that a UPI will be available in all instances (e.g., a patient is unconscious or has not yet been issued a UPI). Given the time-sensitive nature of health care, services will often need to be provided even when the identifier is not readily available. Statistical matching may be more flexible in this regard because of its ability to attempt matching on common personal attributes that may be available from a driver's license or from family members.

One approach to such situations is to issue a temporary identifier (the provider organization could pre-acquire a set of temporary identifiers for emergency use if they wished) and to use it for data linkage until the permanent identifier is determined. The patient's information can then be transferred to the permanent identifier, and the temporary identifier can be retired. But it is likely that, even if a UPI were implemented, a statistical identification method should augment the UPI to achieve robust linking

---

[2] An architecture proposed by Connecting for Health (2005b) to use statistical matching requires a "network of networks." In this scheme, individual health care providers would subscribe to a regional health information network, and each regional network would subscribe to a national network. Each network would have a Record Locator Service (RLS) that links a patient's demographics to locations at which that patient has data. Thus, RLSs would need to be updated as patient demographic data change and new records are added.

under all conditions.[3] One approach would be to include the UPI as the first of many potential matching keys in those systems employing statistical matching. When the UPI is available, the match would be immediate and certain, not requiring the other personal keys. This would allow for gradual implementation of the UPI within existing systems. The UPI might also replace the SSN in those systems that use the SSN, further protecting the identity of the patient.

## Research and Public Health Considerations

By decreasing data-linkage ambiguity and enhancing anonymity, UPIs can facilitate public health and population research in important ways. Both public health and population research often depend on aggregating data from disparate health systems and data sources. Anonymity is often a prerequisite for both; because the UPI supports anonymity, it may streamline these activities.

Further privacy measures can be added to the UPI by linking it to a random ID number used for a specific public health or research analysis, without any personal characteristics being exposed during that analysis. Often in public health and population research analyses, it is important for health care data drawn from different sources to be linked back to an individual to allow the individual to be notified of risk (e.g., if a recently approved medication leads to an enhanced risk for stroke), while keeping the identity of the individual private. This type of linkage requires anonymity and unambiguous matching, both of which are better supported by UPIs.

---

[3] The Kaiser Permanente system of record linking may provide a good example of such a hybrid system. Patients are usually linked with their Kaiser membership number. Without that number, they can be linked on personal attributes, and the system can even use some health conditions to confirm a match.

CHAPTER FIVE

# Privacy and Security of the Alternatives

Much of the controversy swirling around the UPI has little to do with the UPI *per se* and much more to do with what the emerging NHIN architecture, connectivity, and interoperability would mean for privacy. Despite the halt in work on a UPI, HHS continues to develop the NHIN, arguably without adequately addressing the same privacy concerns that caused Congress to table the UPI in the first place.

This chapter briefly discusses some of the technical aspects of privacy and security with a UPI relative to statistical matching and some of the legal considerations related to these methods.

## Privacy and Security of the Primary Alternatives for Patient Identifiers

A comprehensive analysis of options for a unique patient identifier, commissioned by HHS before the congressional ban, supported the use of a UPI and concluded that among its strengths was accurate identification without the "repetitive use and disclosure of an individual's personal identification information," thereby preserving anonymity, protecting privacy, and preventing unauthorized access to health information (Appavu, 1997a). While acknowledging the possibility of risks associated with misuse of the UPI, the author stated, "since access to healthcare information is possible even without the use of a UPI, the solution to this and other legitimate concerns does not lie in eliminating the use of a UPI," and suggested that the threat of "rigorously enforced" legal sanctions would limit the potential for abuse (Appavu, 1997b). Our review of the European experience with UPIs to date revealed no significant breaches of the security of individual health information and only limited concerns about a UPI among patients.

By contrast, proponents of statistical matching suggest that a UPI scheme will reduce privacy by making all of a patient's data recognizable and accessible via the single UPI. *However, if a statistical matching scheme is made as accurate as a UPI, it provides an identical capability to identify and access patient data by using its matching keys.* Furthermore, the matching keys for an algorithm reveal the identity of (and other information about) the patient whose data they identify, whereas a UPI (being just

an alphanumeric value) reveals nothing about the patient. And, in contrast to using personal information, being able to retire a compromised UPI and issue a new replacement UPI should facilitate reestablishing security after a breach of a patient's health information.

It is technically possible to enhance the protection of personal health information with either identifier approach. Simple methods of preventing unauthorized access would probably rely on password protection. Stronger methods would encrypt the identifiers and some of the health information. However, encrypting the personal attributes used in statistical matching may be limited, because it can greatly reduce the power of the statistical matching techniques, which cannot find near matches with encrypted data.

It has been suggested that giving an individual a choice of whether to acquire a UPI would reduce overall privacy concerns. Those worried about misuse of the UPI could simply opt out (Hieb, 2006). From efficiency and operational viewpoints, it would be much more desirable to have a mandatory system in which everyone has a UPI. However, in a voluntary system, many of the benefits of continuity of care and efficiency might be achieved if significant participation of those individuals seeing multiple providers could be obtained, either through physician encouragement or through employer or insurer incentives. Although there is no direct evidence about how many would participate in a voluntary UPI system, a recent poll by the California HealthCare Foundation indicates that 92 percent of patients are willing to share their health information among providers, despite the fact that 67 percent of these same individuals have concerns about the privacy of their records (California HealthCare Foundation, 2005).

The following section discusses the key legal issues associated with the identifier alternative.

## Legal Implications of the Two Main Methods of Patient-Identity Establishment[1]

Under current federal and state privacy rules, the proposed NHIN is likely to generate legal problems regardless of whether it employs a UPI or statistical methods.

In that regard, privacy advocates point to unresolved concerns about the reach of the HIPAA Privacy Rules and the application of the Rules to the emerging NHIN. HIPAA did not anticipate the development of fully interoperable networks, and the Privacy Rules do *not* cover the full panoply of organizations that will be involved in collecting, processing, and using health records in the NHIN. On a somewhat dif-

[1] This is a brief summary of the analysis conducted for this study. For a detailed discussion of the law, our analysis of the merits of the two main approaches to identity establishment, and an enumeration of policy alternatives suggested by various stakeholders in the debate, see Greenberg and Ridgely (2008).

ferent note, since medical providers have a legal obligation to take reasonable steps to ensure that recipients of protected health information do not violate the Privacy Rules, it follows that providers may become risk-averse about participating in future extended networks in which the opportunities for due diligence (e.g., confirming the identity of recipients, the validity of medical-record queries, and/or the security of distant network-access points) are far more limited. Then, too, as we have written about elsewhere (Greenberg and Ridgely, 2008), the scope and application of current HIPAA Privacy Rules (and particularly of the "treatment, payment and operations" safe harbor) are ambiguous with regard to a hypothetical national network: Such ambiguity raises the possibility that some health records will be inadequately protected from disclosures under the law, even as some providers become increasingly risk-averse about participating in the network.

Meanwhile, a range of organizations that participate in future interoperable HIT networks may face increased liability risk for *indexing errors* (i.e., mismatching records and identifier tags) or for other mistakes made in dealing with UPI-tagged records. In addition, state differences in privacy laws, as well as inconsistencies in the federal rules, are likely to create many more barriers for the NHIN to overcome.[2] Inconsistent state privacy laws—particularly with regard to the protections required for sensitive health information, such as that related to mental health, HIV status, and family planning—will likely impede the flow of information across some state lines, limiting the development of a truly national network. Again, none of these concerns is specific to the type of patient identifier utilized in the network.

In contrasting the relative merits of the UPI and statistical methods, we would point out that a unique patient identifier, once developed, would immediately become protected health information under federal and (applicable) state law. UPIs would be sensitive information and could be a target for illicit access. Unlike the demographic components of an algorithm (such as the SSN), however, the UPI would not link to financial records that are the specific target of identity thieves. If the UPI were to facilitate the development of a more fluid national network, it could indirectly have a negative effect on privacy; but that effect could be ameliorated directly through other aspects of systems architecture, such as encryption, access controls, and audit trails. And use of a UPI might actually improve privacy by limiting the transmission of more-sensitive identifiers, such as the combination of names, address, date of birth, and SSN. Our analysis in other sections of this document also suggests that the UPI might facilitate greater interoperability and might also offer benefits in improving the fidelity of records matching.

By contrast, an NHIN that employs statistical matching would involve the massive electronic transmission of existing demographic identifiers (including the SSN) simply for the purpose of locating records. Arguably, the risks of misappropriation are

---

[2] See, generally, the discussion in Greenberg and Ridgely (2008).

much higher in identification schemes that use the SSN; therefore, the use of algorithmic methods potentially poses a more severe risk to privacy interests than does using a UPI. As well, use of statistical matching methods increases the probability of both false-positive and false-negative errors, raising the risk of provider liability for medical errors resulting from imperfect matching.

With regard to the privacy risks associated with a UPI-based NHIN architecture, one avenue that has been suggested for reducing privacy concerns would involve giving each individual consumer a choice of whether actually to acquire a UPI. Those worried about misuse of the UPI could simply opt out, and the federal government could simultaneously enhance patient confidence in the NHIN by creating a network that offers patients a combination of data restrictions and user privileges. We are skeptical about the actual implementation of such an architecture, however. Although it may technically be possible to create this architecture, we doubt that it is likely that millions of Americans, some of them with very limited computer skills and experience, would be able and willing to use a complicated access-control scheme of this kind.

We conclude that the controversy surrounding the UPI is misplaced. The genuine concerns expressed by advocates on behalf of patients have much more to do with the emergence of an NHIN for which existing federal and state legal safeguards are manifestly inadequate. Prohibiting development of the UPI only sidesteps the real problem: that the federal government is proceeding with development of the NHIN without first building for it a firm foundation in privacy law and then building privacy protections directly into the architecture.

CHAPTER SIX

# Costs

In addition to the privacy, security, and legal issues associated with the identifier alternative, it is important to understand the costs of implementing and maintaining the alternative within a national health information network.

The costs of creating and maintaining a patient identifier (the Master Patient Indexes for all the local, regional, and national organizations that maintain them) are only a small portion of the whole cost for a system that also uses the patient ID to make personal health information accessible to health care professionals when and only when they need it. In this chapter, we present only the costs of creating and maintaining a patient identifier. Appendix B provides a discussion of the overall system costs.

The cost of the patient identifier depends on the approach (statistical matching or UPI) and on the architecture chosen to achieve connectivity. To estimate *statistical matching* costs for an NHIN, we used the network-of-networks scheme proposed by Connecting for Health (2005a). In this network, individual health care providers would subscribe to a hierarchical structure of RHIOs and each RHIO would have an RLS that links patient demographics to locations at which that patient has data.

Kaushal et al. (2005), in developing the cost of such an NHIN, estimated the number of RHIOs in the hierarchical organization. In Appendix B, we cite their numbers, the implied sizes of the patient RLS of each RHIO, and approximate cost factors from RLS vendors, to estimate that the total would be a one-time cost of about $90 million and a maintenance cost of about $18 million per year for the necessary RLS to link patients and records nationally with statistical matching.[1]

It is technically feasible to perform statistical matching on providers' medical-record databases without informing patients or seeking their permission to share their data. However, we feel that it is unlikely that such matching would be permitted in a system of national scope. We suppose that enrolling a patient into the system will occur during a visit to his doctor, and will require 5 minutes of the receptionist's time.

[1] RLS vendors typically charge on the order of several hundred thousand dollars to link the records of a million patients spread across multiple data sources, and they charge several million dollars to link the records of tens of millions of patients. A typical maintenance fee is 20 percent of the one-time charge. Personal communication of James Bigelow with MPI vendor (December 2006).

If we value the receptionist's time at $1 per minute, enrolling 300 million people will cost $1.5 billion. If a patient must be re-enrolled annually, this cost will be incurred again each year.

A *voluntary UPI system* as proposed by the ASTM would have an estimated cost of $25 million for the first five years for the national organization issuing UPIs (Hieb, 2006). As before, patients must be enrolled annually, at a cost of $1.5 billion per year. This UPI is not canonical, and an individual might obtain multiple UPIs by registering in different RHIOs. Doing so can, in turn, create false-negative errors if the patient does not present all of his or her UPIs. This approach, being voluntary, requires, for an NHIN, that a hierarchical RHIO structure and statistical matching be developed in parallel for those patients who opt out of the UPI, thus incurring the additional costs of the statistical matching process as well.

*Mandatory UPIs* that are guaranteed to be unique and canonical for everyone in the United States would require a central administrative organization. They would also require an accurate authentication process to ensure that the person being registered is who he or she claims to be. Here, we based our figures on several existing estimates. The SSA (1997) estimated that issuing an "enhanced" SSN—one that has improved security features and also satisfies most of the ASTM criteria—for 277 million current SSN holders would cost between $3.9 and $9.2 billion, depending on the security features built into the new card. The National Governors Association (2006) has estimated that issuing a "Real ID" (i.e., one ID per person and one person per ID), based on the Real ID Act for establishing an authenticated identifier through the issuance of state driver's licenses, would cost about $37 per ID. The total cost of issuance ($11.1 billion for 300 million individuals) is consistent with the upper end of the SSN estimate.[2]

To put these costs in perspective, previous studies of the value of connected EHR systems estimated a potential efficiency savings of $77 billion per year at the 90-percent level of adoption; added value for safety and health could double these savings (Girosi, Meili, and Scoville, 2005).

A one-time cost of $1.5 to $11.1 billion for a UPI, to remove the systemic errors in health-records retrieval, is small by comparison. A significant reduction in duplicate testing and imaging that could result from more-complete retrieval of medical records could potentially save $4 billion per year (Girosi, Meili, and Scoville, 2005). Avoiding adverse drug events, which are often the result of incomplete linking to information about a patient's medications or allergies, could save an additional $4.5 billion per year (Bigelow et al., 2005). And, to the extent that lack of a UPI prevents the health care system from achieving full automation and connectivity, and its full clinical, public

---

[2] These estimates are based on the assumption that federal or state infrastructures exist to administer these programs, and that if agencies such as the SSA, state motor vehicles departments, or county health agencies cannot take on this function, costs may increase.

health, and research potential because it is dependent on nonambiguous linking of patients and records, the value of a UPI may be considerably higher than the possible $11.1 billion cost.

CHAPTER SEVEN

# The Policy and Political Environment for a National Patient Identifier

In addition to the technical merits, legal issues, and costs of the alternative identifier systems, it is important to understand the policy environment.

HIPAA combined the promotion of standards-based HIT-enabled efficiencies in health care administration with national standards for maintaining the privacy and security of patient information. But lack of confidence in the privacy and security systems, practices, and regulations in the late 1990s caused Congress to subsequently stop HHS from developing the UPI systems mandated in HIPAA. Since that time, many have questioned whether a UPI system, or any national patient-identification and information-networking system, could gain support from Congress and the American public. A full analysis of politics and policy options in these areas was beyond the scope of this project. However, in addition to our technical comparison of UPI and statistical matching systems, we briefly scanned the political and policy environment in which this study's recommendations would be received.

## The Politics of Privacy and Security

The key political consideration for developing national patient-identification policy today, as in the late 1990s, is widespread public concern for privacy and security. One survey, for example, found 67 percent of Americans expressing concern about the privacy of their personal medical records (California HealthCare Foundation, 2005). Over half of those surveyed (52 percent) are concerned that their personal information might be seen by their employer and used to limit job opportunities. That same survey found that 24 percent are aware of specific breaches in security; and of those, 66 percent are more concerned about medical-record privacy as a result. Overall, 72 percent of those surveyed believe computers increase the risk of their data being broken into.

In a Markle Foundation (2006) survey, 80 percent of Americans said that they are very concerned about identity theft or fraud; 77 percent are very concerned about marketers gaining access to their data; 56 percent, about employers gaining access; and 53 percent, about insurance companies. A Harris Interactive (2007a) study found that half of adults (50 percent) believe that patients have lost control over how their medi-

cal records are used by organizations, such as life insurers, employers, and government health agencies. These concerns color attitudes toward electronic networking in both non-health and health care–related applications.

But in some ways, breaches of health care information are more troubling. Unlike money stolen from a bank account, health information cannot be put back into the vault once it is out; and the damage done may be both more difficult to assess and more permanent. When developing HIT policy, Congress cannot ignore such widespread concern.

Policymakers' efforts to address Americans' privacy and security concerns are complicated by the lack of general agreement on the meaning of the term *privacy and security* or on the rights, obligations, controls, and infrastructure implied by endorsing them as policy goals.[1] In addition, privacy and consumer groups vary in regard to their attitudes toward patient-identification systems and health-information networking. Some libertarian and citizen's-rights groups object to the development of any national ID system, perceiving them as a threat to their assumed right to autonomy and/or out of concern that the national database of information on individual citizens required to create an identification system could lead to unwarranted surveillance, harassment, discrimination, and identity theft.[2] These were among the reasons, for example, that the American Civil Liberties Union (ACLU) opposed a national ID card and the Real ID Act of 2005.[3]

Yet many other interest groups have also publicly recognized the value of converting from paper to electronic medical records and of using networks to share authorized patient information for improving health care.[4] A number of coalitions representing these groups, HIT and health care industry groups, and others have attempted to produce and promote principles for ensuring privacy and security in electronic patient-information exchange. A 2007 e-Health Initiative publication summarized a number of these efforts, including efforts of the American Health Information Community (AHIC), American Medical Informatics Association (AMIA), Connecting for Health–Markle Foundation, and others (e-Health Initiative, 2007). A statement of principles cited in that publication and developed by the Coalition for Patient Privacy (2007) does a good job of summarizing the policy directions sought by many privacy advocates.

---

[1] For definitional discussions on privacy, confidentiality, and security, see ASTM (2004); and Institute of Medicine (IOM), Committee on the Disposition of the Air Force Health Study (2006).

[2] The ASTM approach does not utilize a centralized national database, but instead utilizes regional, or distributed, databases, as does an NHIN composed of linked RHIOs.

[3] For a discussion of these and related concerns, see Stanley and Steinhardt (2003).

[4] For example, the Coalition for Patient Privacy includes 47 consumer, provider, and privacy and patient advocate organizations that support EMR adoption and networking with proper privacy and security controls. See the press release, including a Coalition membership list, at the Patient Privacy Rights website (no date).

Although these principles do not specifically mention the UPI or patient-identification standards, privacy groups have also long valued the use of codes to mask patient names in research and medical surveillance. For example, ACLU (1999) argued in favor of the use of unique patient identifiers instead of patient names in the case of the Centers for Disease Control and Prevention's (CDC's) proposed guidelines for HIV-monitoring programs, stating that "a unique identifier system has proved remarkably effective in Maryland—the one state that has made a serious attempt at HIV surveillance without names."

Today, privacy and consumer groups have a much greater appreciation of the value of EMRs and networking than in the late 1990s, when the UPI was initially being debated; and there is greater understanding of the technologies and policy options now available that could ensure reasonable privacy and security for patient information in a networked environment. In a 2006 letter to Congress, for example, the Coalition for Patient Privacy applauded Congress's "efforts toward bringing the American healthcare system into the 21st Century by using technology to control costs and to reduce medical errors" (Coalition for Patient Privacy, 2006). But it qualified this support with a call for Congress "to build a patient-centered system with patient privacy rights as the core of the health IT system." This qualification reflects the general view among privacy advocates that HIPAA sets too low a floor for national privacy protection and that HHS's leadership in protecting American's privacy has been weak.

This low opinion is reflected in a 2005 survey that found that only about one-quarter (27 percent) of Americans believe they have more rights as a result of HIPAA and only 20 percent are willing to share their health data with government (a possible indicator of lack of trust) (California HealthCare Foundation, 2005). It is also reflected in a 2007 Harris Interactive survey conducted for IOM, reportedly showing that nearly three out of five Americans agree that the privacy of their health information is not well protected by federal and state laws and organizational practices (Ferris, 2007). Finally, a 2007 Government Accountability Office (GAO) report was critical of HHS for failing to develop a coherent strategy and integrated approach to protecting privacy. These findings suggest that privacy and consumer-rights groups have both a basis for their concern and significant public support behind their call for change. For a more complete discussion of this topic, and of potential changes in federal and state policy to address them, see the legal analysis resulting from this study by Greenberg and Ridgely (2008).

Despite these general concerns for privacy and security, a good deal of popular support has been expressed for selected HIT and networking applications. Polls show that 98 percent of patients are willing to share information with their own doctor, and 93 percent believe that computer-based systems give doctors quicker access to data (California HealthCare Foundation, 2005). A more recent poll (Harris Interactive, 2007b) shows that 70 percent of U.S. adults agree that they are generally satisfied with the way doctors and hospitals handle protecting the confidentiality and security of

personal health: 20 percent strongly agree with this, and 50 percent somewhat agree. And, by 63 percent to 25 percent, a majority agrees that increased use of computers to record and share patient medical records can be accomplished without jeopardizing proper patient privacy rights.

Another poll found that 65 percent of the public is interested in accessing their own personal health information electronically, and 82 percent would be interested in keeping track of their children's health records and services, such as immunization dates.[5] This poll reports that Americans believe they could gain more control over their health care by using electronic personal health records, and 90 percent said that it would be personally important to track their symptoms or changes in health care online. Nearly nine in 10 Americans (88 percent) say online records would be important in reducing the number of unnecessary or repeated tests and procedures they undergo. More than eight in 10 (84 percent) would be interested in accessing their electronic records to check for mistakes—even higher proportions of African Americans and Latinos express this interest. And, four-fifths say they would be interested in managing the financial aspects of their health care, such as tracking insurance payments and out-of-pocket costs online. Yet the value of this support is likely conditional on a public sense that privacy and security are adequately protected, especially given the high levels of concern about identity theft or fraud found among those in fair or poor health (Lake Research Partners and American Viewpoint, 2006).

For the purpose of this study, however, it is important to note that the vast majority of individuals and organizations weighing in on this issue are focused on privacy and security concerns, not on patient-identification policy. Except for those opposed to any national patient-identification system or standard, very few organizations have addressed the distinctions between statistical matching and the UPI. If they were publicly debated, it is likely that the added privacy and security risks associated with statistical matching would become an issue. But without this debate, statistical matching has had the advantage of not requiring new national policy and has therefore avoided being judged under the bright lights of public scrutiny.

## Striving for Health Care Quality and Efficiency

Accurate and unambiguous patient identification is important in providing clinical care, managing patient records, and supporting clinicians' administrative and financial systems. In addition, the interoperability of EMR systems and the efficient exchange of data between systems is dependent on effective patient-identification management.

---

[5] Lake Research Partners (LRP) and American Viewpoint's National Survey on Electronic Personal Health Records, November 2006. This survey was done in preparation for the Markle Foundation's conference, *Connecting Americans to Their Health Care: Empowered Consumers, Personal Health Records and Emerging Technologies*, December 7–8, 2006, Washington, D.C.

Hillestad et al. (2005) showed that nearly full adoption of interoperable EMR systems in the United States could save $81 billion annually by improving health care efficiency and safety. Further, HIT-enabled improvements in prevention and management of chronic disease could eventually double those savings while lowering age-adjusted mortality 18 percent by eliminating 404,000 unnecessary deaths and reducing annual employee sick days by 40 million; transaction efficiencies could add another $10 billion or more in annual savings. Bower (2005) suggests that these potential savings could more than double again—to $346 billion a year or more—if health care were sufficiently transformed to generate the relatively modest 1.5-percent annual productivity gains that were realized from IT-enabled efficiencies in the retail and wholesale industries. These are tremendous benefits that depend on accurate patient identification and interoperability. But surprisingly, these topics almost never come up during policy debates on national patient-identification policy.

What does come up frequently, however, is providers' concern that, without accurate patient-identification systems, they will encounter false-positive and false-negative matching when networking patient data. In a local hospital or integrated delivery network (IDN) setting, providers have been able to help guard against the negative consequences of matching errors by using manual review of information and cross-checking against actual patient-generated data or physical findings.

But even in those settings, many providers are uncomfortable with the random nature of statistical matching. This concern was a big part of why the American Hospital Association (AHA) supports the development of a UPI, stating, "without a single authentication number, there are serious safety risks that could arise from attributing a medical record to the wrong individual . . . a cluster of demographic information may not be sufficient to distinguish between the 37-year-old Mary Jones with diabetes and a penicillin allergy and the 37-year-old Mary Jones in perfect health. Mixing up their records could have serious consequences" (AHA, 2006).

As demonstrated in Chapter Three, statistical matching across multiple disparate delivery systems, RHIOs, and the NHIN can carry even greater risk of error. The AHA's position is that "the electronic exchange of health information requires a consistent, reliable mechanism for matching patients to their records. This is best achieved with an individual health information authentication number" (AHA, 2006).

Many other national organizations have joined this call for federal leadership in the development of a UPI or alternate national solution to a unique patient identifier. A 2003 IOM report, focusing on the critical role of HIT in clinical care and quality improvement, called on Congress and the Executive branch to take steps toward a proactive solution to the unique patient identifier issue (IOM, 2003). A Healthcare Information Management and Systems Society (HIMSS, 2003) position statement that same year declared that, without a patient identifier, whether unique or voluntary, true data interoperability is not possible.

The Alliance (2007), representing a broad range of senior health care executives, has also called for development of a voluntary, unique patient identifier. American Health Insurance Plans (AHIP) supports the idea that "a uniform method for identifying individual consumers needs to be established," and that "HHS should evaluate whether to implement a universal consumer identifier."[6] The American College of Physicians' (ACP's) position states,

> ACP believes that there are patient safety benefits in the use of a unique patient identifier that far [outweigh] any reasonable privacy or government intrusion concerns. The College recommends that HHS use its resources to place this issue "on the table" for further discussion" (Stubbs, 2005).

While most health care provider organizations see the unambiguous identification of patients as critical to locating and accessing patient information electronically, a few provider organizations have specifically opposed the UPI. The American Psychoanalytic Association (1999), for example, "strongly opposes the development and use of a unique, universal health identifier on a national or regional basis, as well as the mandatory reporting of treatment contacts (encounters) to a national, regional, or local health care databases." This does not imply that mental health care specialists oppose electronic medical records or controlled networking of patient information among local providers directly involved in the care of their patients. The position is likely a reflection of the unique nature of mental health diagnosis and care and the trust and confidentiality required between its practitioners and patients. But such trust is an important component of every provider-patient encounter, and the lack of trust creates dangerous barriers to frank and honest communication in every specialty.

In a 2005 poll (California HealthCare Foundation, 2005), 12.5 percent of Americans reported hiding information from their medical record. This poll was unrelated to the use of EMRs. A later survey (Lake Research Partners and American Viewpoint, 2006) found that, although almost eight in ten (77 percent) adults in the United States say that they have not withheld information, a significant one in six (17 percent) say that they have withheld information. Even more worrisome from a health care perspective, the percentage of withholding information rose to over one in five (21 percent) among those who are in only fair or poor health.

These results may say as much about the perverse incentives in America's health insurance system as they do about privacy concerns; however, they reinforce the need to couple strong assurances of privacy and security control with any effort to expand the networking of health care data. A national initiative to promote the NHIN and electronically identify patients would be of limited value if a high percentage of patients

---

[6] Susan Piasane, Vice President of Communications, AHIP, personal communication with Roger Taylor, August 19, 2008.

and providers opted out or if an increasing percentage of patients began withholding vital information from their health care providers.

Two of the major value-producing areas for HIT identified in Hillestad et al. (2005) were chronic disease management and increased consumer involvement in managing their health care, including preventive care. The AHIC (2006) recognized these same areas when it made consumer empowerment and chronic care the first two of its four 2006 breakthrough focus areas. AHIC defined consumer empowerment as follows:

> Make available a consumer-directed and secure electronic record of health care registration information and a medication history for patients.

The chronic care need was described as follows:

> Allow the widespread use of secure messaging, as appropriate, as a means of communication between doctors and patients about care delivery.

HIT-enabled chronic disease management has the potential to significantly improve health and health care while improving efficiency. And this potential will grow over time. It is estimated that, by 2010, 40 percent of Americans will suffer from a chronic disease (Cain, Sarasohn-Kahn, and Wayne, 2000). Chronically ill consumers are interested in taking a more active and informed role in their heath care management, and they can benefit from the personalization and access to tools available from Web-based technology (Row and Metzger, 2001). Using e-disease management programs, clinicians, case managers, and patients link to each other and to patients' data through direct dial-up, local area networks, and/or through Internet service providers to Web-based programs (Row and Metzger, 2001). Each of these applications, however, requires unambiguous patient identification and interoperable records.

Finally, controlled networking of selected patient-specific clinical information could also significantly improve the efficiency and effectiveness of medical research and public health initiatives. Although a review of these HIT applications and their potential benefits is beyond the scope of this study, it is important to note that Americans may be willing to support such broader networking of data in those situations in which they can see that a valid reason exists—and as long as their identity is protected.

In one study (Markle Foundation, 2006), those valid reasons included sharing information with public health officials to detect disease outbreaks (73 percent) or bioterrorist attacks (58 percent); with researchers, doctors, and hospitals to learn how to improve quality of care (72 percent); and with appropriate officials to detect medical fraud (71 percent). However, the survey found that, when asked, most Americans say they want to have some control over the use of their information for such purposes. To gain the needed public trust, Congress would be wise to integrate additional privacy and security controls into any legislative effort to promote the NHIN or to authorize

the sharing of personal patient care information with the research or public health communities.

## Integration with Existing Standards

The lack of federal leadership in establishing patient-identifier standards and in funding projects to test or develop a UPI has left the health and HIT industry to define their own paths. As discussed above, ASTM (2000) established standards for the ideal properties of a unique identifier. HIMSS proposed, and the ASTM Committee E31 on Healthcare Informatics has developed draft standards for, a voluntary national health care identification system (ASTM, 2006). Connecting for Health studied and recommended a statistical matching system rather than the more contentious UPI, and both their Health's Technical Guides (no date) and Integrating the Healthcare Enterprise's (IHE's) Technical Frameworks (various dates) utilize this statistical method.

HHS's adopted HIPAA eligibility (X12 270/271) transaction standards allow individuals to be identified with both demographic data (e.g., first and last name, date of birth) and the health plan's unique ID number for that subscriber (subscriber's primary ID number).[7] The Council for Affordable Quality Healthcare (CAQH), a collaboration of the nation's leading health plans, networks, and industry trade associations,[8] is incorporating these standards into operating rules for eligibility and benefits transactions.

However, none of the standards constitutes a true national or industry standard. Without national standards, the marketplace has generally adopted the expedient approach of using the same statistical matching methods that most large health care provider systems have used in their MPI systems for years. Likewise, Regional Health Information (Exchange) Organizations and the consortiums developing NHIN models under contract with HHS have largely limited themselves to various forms of statistical matching, despite its known limitations in linking records from large, diverse databases. The resulting default standard practice, absent a UPI option, has been statistical matching.

To date, however, evolving U.S. industry standards have not endorsed statistical matching to the exclusion of a UPI. For example, Health Level Seven's (HL7, 2006)

---

[7] Note that, for eligibility inquiries and responses, ANSI ASC X12N 270/271 Version 4010A implementation guide (*Benefit Eligibility Inquiry/Response Transactions: Washington State Medical Assistance Administration Companion Guide* [ACS EDI Gateway, Inc., 2005] established a standard for compliance with HIPAA's Electronic Data Interchange (EDI) standards. A new version is in development and some provider organizations have asked that the individual's health plan number not be required. Insurers and employers have reportedly resisted that change because of privacy and accuracy concerns associated with probabilistic matching without the use of any unique identifiers.

[8] For more information, see the Council for Affordable Quality Healthcare homepage, no date.

Context Management (Clinical Context Object Workgroup [CCOW]) Standards, designed to enable secure linkages and integration between applications, include as patient/subject identifiers both a patient's national identifier number and a list of other potential patient identifiers.

As the source of standards for personal identity, The Health Information Technology Standards Panel (HITSP)—a public–private sector partnership administered by the American National Standards Institute and under contract with HHS—does not specifically include a UPI in its evolving standards. It utilizes IHE's frameworks, employing available demographic identifiers and statistical matching methods to find patients across disparate identifier domains and to cross-reference these identifiers in a Patient Identifier Cross-Reference Manager and/or Master Patient Index.[9] But there is no prohibition against using a UPI, if it were available, as one of the patient identifiers or demographic data elements. Likewise, the requirements for EHR certification developed by the Certification Commission for Health Information Technology (CCHIT), a consortium under HHS contract, seem to be flexible enough to allow use of a UPI. These standards simply require that EHRs include, as a minimum, the ability to store key patient identifiers and demographic information and to look up patients' records using this information (CCHIT, 2007).

Retaining the potential to use UPI as a key demographic data element to identify patient records is important, both for international compatibility and because a UPI may eventually be seen as necessary in the United States. Further, allowing the addition of an optional UPI as one of many search keys could greatly reduce the chance of erroneous matches for those patients choosing to have a UPI.

But policymakers should realize that the systems designed to ensure privacy, security, and search efficiency in UPI-based information-exchange networks are quite different from those designed for statistical matching–based networks. Not only is it easier to assure patients that they will have authorization control over all personal identifying information in a UPI-based system, but the searching and matching processes are more efficient and the need for RHIOs or other hierarchical hubs to manage RLSs are significantly reduced—an important distinction, given concerns about the financial viability of RHIOs. Many of the efficiency advantages of security, privacy, structural, and search features that could be built into a UPI-based system would be lost if a UPI was only later added as one of many identifying data elements for statistical matching.

Finally, it should be acknowledged that this monograph largely assumes that most clinical electronic patient-care information will reside in provider-based EMRs, and the sharing of patient information between their EMRs and other health care entities will go through a health information exchange network. Standards exist or are in development for most aspects of this model, and the model supports a wide variety of

---

[9] For more information, see HITSP (2007).

electronic transactions involving patient information, including administrative transactions; access to patient-condition-specific reference material; the exchange of patient-specific data elements; the transfer of various forms of record summaries, such as the Continuity of Care Record (CCR); and the transfer of complete EMRs.

But standards are under discussion for other models of information storage and exchange as well: various models of Personal Health Records (PHRs), which could be, for example, stored in the patient's personal computer, domiciled in a free-standing service provider or health data bank, or tethered to some other organizational relationship with the patient, such as the patient's insurer, employer, or health care provider. Some of these models have the ability to extensively utilize their own entity-specific information database and patient-identity system. For example, insurers and employers may have selected claims data and unique alphanumeric identifiers for each covered individual and employee. Yet most individual patients will have multiple employers and insurers over time; and many will have multiple providers at any one point in time. Keeping their PHR up to date and useful will likely involve multiple exchanges of data between disparate entities with a variety of approaches to patient identification.

Except for those few PHR models that have an individual or their agent enter, update, and transmit their needed health care data personally, most successful PHR systems will likely require standard networking capabilities to be effective. Therefore, most PHRs will need to be able to exchange data through networks using either statistical matching or a UPI, much in the same way as EMRs. Using this assumption, we did not separately analyze alternate patient-identification options for PHRs.

CHAPTER EIGHT

# Conclusions and Implications

## Conclusions

### Broad Adoption of a UPI Should Enhance the U.S. Health Care System

The foregoing analysis indicates that a health care system in which every patient has a unique, nondisclosing patient identifier is clearly desirable for reducing errors, simplifying interoperability, promoting NHIN architectural flexibility, and protecting patient privacy.

### A Hybrid System Utilizing Both Statistical Matching and a UPI Will Be Necessary for the Foreseeable Future

A hybrid system will be necessary when only some of the population has a UPI (as in a voluntary system), during the implementation of a UPI (which may take a number of years), and when a patient cannot provide his or her UPI and health services must be rendered. In contrast, depending only on statistical matching will perpetuate errors in health-records retrieval because of statistical matching's reliance on nonpermanent and non-unique personal attributes. One approach to phasing in a UPI is to use it as an additional attribute in statistical matching. When a UPI is available, the match should be immediate and certain. When it is not, the matching process will be no worse than that of the present system.

### Security and Privacy Could Be Strengthened with a UPI

In the context of a networked health information system, security and privacy have much more to do with how access is managed and records maintained than with a specific identifier approach. Password protection and encryption of a UPI are relatively easy, whereas encryption of personal keys used in matching algorithms decreases the power of the algorithms. Repeated disclosure of personal information and linking personal information to health information, required in statistical matching in a network, probably carries a greater security risk of disclosure of sensitive information than a UPI.

We conclude that the controversy surrounding the UPI and privacy is misplaced. The genuine concerns expressed by advocates on behalf of patients have much more

to do with the emergence of an NHIN for which existing federal and state legal safeguards are manifestly inadequate. Prohibiting development of the UPI only sidesteps the real problem—that the federal government is proceeding with development of the NHIN without first building for it a firm foundation in privacy law and then building privacy protections directly into the architecture.

### Costs of a UPI Are Significant, but Probably Much Less Than the Value Associated with Error Reduction, Efficiency, and Interconnectivity of the Health Care System

Costs depend on the scope of uniqueness and how strong and centrally managed the registration and authentication processes are. Guaranteeing that everyone in the United States has a health identifier that is unique and canonical would likely cost about the same as the proposed Real ID system (approximately $11.1 billion for one-time issuance) and would require a national infrastructure for support. Issuing and managing UPIs that are regionally unique but not canonical would cost about $1.5 billion. A system based solely on statistical matching would cost about $90 million initially, and it would incur a maintenance cost of about $18 million per year for the identifier portion of the system, assuming that an effective, hierarchical NHIN infrastructure exists throughout the country.

## Implications for Public Policy

HHS has not funded any development work on the UPI since the late 1990s. Consequently, none of the NHIN "consortium" contracts funded by HHS has employed a UPI approach, which limits a key purpose of the consortium process: to experiment with and develop the best approaches to interconnectivity and interoperability. Privacy and security are clearly inadequate under current law and must be enhanced as the health care system becomes digitized and interconnected. However, prohibiting development of a UPI actually sidesteps the larger problem: the development of an NHIN without first establishing a legal environment that best protects privacy while also encouraging the advances that interoperability would bring to health care quality and efficiency.

Although it is beyond the scope of this monograph to suggest specific policy actions the government might take with respect to privacy, access, and security of health care information, it is within its scope to recommend that Congress remove the current and clearly counterproductive constraints on HHS with regard to the UPI. Instead, Congress should be encouraging HHS to make a full assessment of the privacy, security, and operational implications of all of the alternatives for linking patients to their health records within the NHIN. These issues should be the subject of open study and debate in the vitally important process of developing the best interoperable U.S. health care system and reducing the errors and inefficiencies in that system. Con-

tinuing *de facto* endorsement of the statistical matching method as the only practicable approach to linking patients to their electronic health records is likely to inhibit the effective development of the NHIN.

APPENDIX A

# Analysis of False-Positive Errors in a Large Demographic Database

The Social Security Death Master File (SSDMF, updated quarterly) contains the name, Social Security number (SSN), dates of birth and death, and last known zip codes for everyone who has registered with the Social Security Administration and has died. The entries all have distinct SSNs, but, as noted in the main text, may not correspond to distinct people.

The data fields used in this study are first name and last name, birth date (and birth year), zip code (two zip codes are given, last known address and where the final lump-sum payment is mailed: the former was used except when blank, in which case the zip code of the final lump-sum payment is used), and we also include the last four digits of the SSN as a data field. The SSN is the only uniformly randomly distributed data field in the set that is completely uncorrelated to any of the other data fields.

Given a subset of these data fields (which we refer to as *database keys*), we were interested in what fraction of the entries of the database (or a subset of the database) matched exactly $k$ entries in the database on all of the database keys (for $k = 1, 2, 3, \ldots, n$). Here, *exact match* means precisely what is described: No processing was done on the data (e.g., removal of extraneous spaces), so "MC CORQUODALE" does not match "MCCORQUODALE" or "MCCORQUODALE JR." This probably results in a negligibly smaller match, or collision, rate than the true collision rate, but the effort to root out all of the quirks in the data fields of the SSDMF would have probably exceeded that of the analysis itself.

Because the database is so large, comparing every entry with every other entry would be an incredibly onerous task. So, before doing any analysis on the database, we separated the database into blocks by the first three letters of the last name. Since last name was always used as a database key, the separation guaranteed that two entries that would potentially match would reside in the same block of data. For particularly large blocks (e.g., last names beginning with SMI and JON), we separated them further by the fourth letter of the last name.

Some definitions and notation are needed. A *collision group* is a maximal collection of database entries that agrees on some set of database keys. For a set of database keys, let $C_i$ be the number of people who are in a collision group of size $i$. To clarify, a database entry that matches with no other entry except itself is in a collision group

of size one. Accordingly, we can define the *collision distribution* $\left(C_1, C_2, \ldots, C_M\right)$ for a set of database keys, where $M$ is the size of the largest collision group. Note that $P = C_1 + C_2 + \cdots + C_M$ is equal to the size of the database under consideration, since every entry appears in exactly one collision group.

For a random entry in the database, the probability of its not colliding with any other entry in the database is simply $C_1 / P$, which means that the *collision rate* (that is, the frequency at which randomly chosen entries in the database will report false positives using that set of keys) is

$$\frac{1}{1 - \Pr(0 \text{ collisions})} = \frac{1}{1 - \frac{C_1}{P}} = \frac{P}{P - C_1}.$$

Similarly, for a random entry, the expected number of collisions with other entries in the database is

$$\frac{\sum_{k=1}^{M} (k-1) C_k}{P}.$$

We can also consider how the collision rate drops for subsets of the population. Given a collision distribution $\left(C_1, C_2, \ldots, C_M\right)$ on a database of size $P$, if we pick a random subset of size $sP$ ($1 < s < 0$), then for a randomly selected individual the probability of zero collisions is

$$\Pr(0 \text{ collisions}) = \sum_{k=1}^{M} \left(\frac{C_k}{P}\right)\left(1 - s\right)^{k-1}.$$

To see this, note that the first term in the summand is the probability that an individual is in a collision group of size $k$. The second term is the probability that, given that an entry is in a collision group of size $k$, none of the other members of the collision group is included in the subset. A similar analysis could be done to develop the rest of the collision distribution for the smaller subset of the database, but our primary interest was the probability of colliding with no other entries.

To extrapolate to larger databases, we made the assumption that, in the new population of size $P + N$, there are no new noncolliding members (that is, every new person will collide with someone already in the population). So, the question becomes what

fraction of the already-noncolliding ($C_1/P$) will still be noncolliding with the addition of $N$ new members?

We turn the problem into a variant of the "coupon collector's problem" (Von Schelling, 1954). The problem is as follows: Given an urn with $C_1$ red balls and $P - C_1$ blue balls, we repeat the following process $N$ times. If the ball is red, we paint it blue and replace it in the urn. Draw a ball; if it is blue, we simply replace it in the urn. A red ball corresponds to a noncolliding member of the current population. Each of the $N$ balls drawn corresponds to a new person being added to the population (and, therefore, to some match, or collision, group). If the ball drawn is red, that person is no longer noncolliding, so the ball becomes blue.

Let $E(C_1,P,N)$ be the expected number of red balls remaining after $N$ draws using this process (where there are $C_1$ red balls initially and $P$ balls total). Let

$$R_{C_1,P}(m,d)$$

be the probability that $m$ balls are painted red after $d$ draws. Then the values of $R_{C_1,P}(m,d)$ can be calculated recursively:

$$R_{C_1,P}(0,0) = 1 \text{ and } R_{C_1,P}(m,d) = R_{C_1,P}(m,d-1)\frac{(P-C_1)+d}{P}$$
$$+R_{C_1,P}(m-1,d-1)\frac{C_1-(d-1)}{P}.$$

Then

$$E\left(C_1,P,N\right) = \sum_{k=0}^{R} kR_{C_1,P}(C_1-k,N) = C_1\left(1-\frac{1}{P}\right)^N \approx C_1 e^{-P/N}.$$

The last equality uses the linearity of expectation on the $C_1$ red balls, and the approximation only holds for large values of $P$ and $N$.

We acknowledge that there are two major issues with this estimation: (1) It assumes that all potential sets of database keys already appear at least once in the database (that is, no new sets of database keys will be introduced into the data set) and (2) that all sets of database keys are equally likely. We know that first and last names, for example, are neither uniformly distributed nor uncorrelated. Doing a similar coupon collector's–like analysis for unequal probabilities becomes considerably more difficult. Instead of drawing from an already-existing population, we could also phrase the question in the following manner: Given a collection of $K$ unique sets of database keys (all

equally likely), the probability of an individual in a population of size $P$ being in a collision group of size $C$ is

$$\binom{P}{C-1}\left(\frac{1}{K}\right)^{C-1}\left(1-\frac{1}{K}\right)^{P-C}.$$

APPENDIX B

# Designs and Costs of Systems for Accessing Personal Health Information

In this appendix, we ask what it would cost to make personal health information (PHI)[1] more accessible to those in the health care system who could use it to the benefit of patients. True, we are primarily interested in mechanisms for identifying patients, but the mechanism one elects to use and the cost of setting it up and using it will depend on its context—i.e., the system in which it is embedded.

In today's health care system, almost all PHI is generated in the course of encounters with health care providers, and a provider not involved in the encounter has little chance of obtaining that information. For example, your primary care physician may write a prescription for you, but if you see another physician for some reason, he will generally not know about that prescription unless you think to tell him. Thus, at present, PHI resides in numerous data systems, including medical records maintained by provider organizations (physician groups, hospitals, etc.) and administrative data (claims, eligibility, and benefits information) maintained by payer organizations.

In this appendix, then, we ask what it would cost to make this PHI accessible to the second physician in our example. We consider three approaches. The first gives health care providers access to information held by other providers. The second gives providers access to information held by payers. The third populates Personal Health Records (PHRs) with information from providers and/or payers, and allows each patient to grant access to any provider he sees. We chose these systems to represent the range of examples we found in the literature and to rule out our considering a particular system as an endorsement.

Four roles must be performed in each system. First, someone must provide the PHI (the *source*). Next, someone must request a patient's PHI (the *user*). Third, someone must store the PHI that the user requests (the *host*). Finally, someone must mediate the transaction (the *broker*). A single party, of course, may play several roles.

We limit our attention to PHI stored electronically and accessed via the Internet.

---

[1] PHI is not the same as a Personal Health Record (PHR), although some of the information considered PHI would certainly reside in a PHR.

## Preliminaries

Even before we discuss the designs of our systems, we can anticipate that some of the components of cost will take the form of (number of players) × (cost per player), where providers, payers, and patients are among the potential *players*. Table B.1 shows recent player inventories.

Equally, we can anticipate that some cost components will take the form of (number of transactions) × (cost per transaction). Visits to physicians' offices and clinics,

**Table B.1**
**Inventories of Providers, Payers, and Patients**

| Entity | Inventory | Source |
|---|---|---|
| Physician Offices and Clinics | | |
| Practices in 2002 | 203,118 | EC02 |
| Office-based physicians in 2004 | 538,538 | HUS06 |
| Hospitals | | |
| Number in 2004 | 5,759 | HUS06 |
| Beds in 2004 | 955,768 | HUS06 |
| Nursing Homes | | |
| Number in 2004 | 16,117 | HUS06 |
| Beds in 2004 | 1,744,258 | HUS06 |
| Residents in 2004 | 1,442,503 | HUS06 |
| Prescription-drug outlets | | |
| Pharmacies and drugstores in 2002 | 39,121 | EC02 |
| Supermarkets in 2002 | 19,721 | EC02 |
| Warehouse/discount dept stores in 2002 | 7,440 | EC02 |
| All other in 2002 | 1,606 | EC02 |
| Miscellaneous | | |
| Home health agencies in 2004 | 7,519 | HUS06 |
| Medical laboratories in 2002 | 5,510 | EC02 |
| Diagnostic imaging centers in 2002 | 5,569 | EC02 |
| Direct Health and Medical Insurance Carriers | | |
| Firms in 2002 | 908 | EC02 |
| Establishments in 2002 | 4,415 | EC02 |
| Population (patients) | | |
| Number in 2004 | 293,655,000 | HUS06 |

SOURCES: EC02 – U.S. Census Bureau, 2002 Economic Census, 2002; HUS06 – National Center for Health Statistics, *Health United States, 2006*, Hyattsville, Md., 2006, Table 117.

hospital inpatient stays, and filled prescriptions are examples of transactions. Table B.2 shows recent inventories of transactions.

### A Note on Security

We divided security measures into two parts: (a) those that protect the network through which users of the system access PHI and (b) those that ensure that the patient ID is not compromised.

We assumed that the networks and applications in our systems have the standard security features that current Electronic Medical Records (EMRs) typically possess—a firewall to prevent hackers from stealing data and a username/password facility to keep track of which insiders are accessing the system—and that the costs for such features are included in the estimates we cite. Of course, the features must be used properly to provide security, and often they are not ("More UCLA Medical Center Employees Peeked at Celebrities' Records, State Says," 2008). Spending yet more money on more-sophisticated security technologies will not improve security, unless people will use them.

**Table B.2**
**Factors Affecting Volume of Transactions**

| Factor | Size or Number | Source |
|---|---|---|
| Physician Offices and Clinics | | |
| Visits to physician offices in 2004 | 990,404,584 | MEPS04 |
| Hospitals | | |
| Inpatient stays in 2004 | 30,194,246 | MEPS04 |
| Outpatient visits to MD in 2004 | 53,120,952 | MEPS04 |
| ED visits in 2004 | 53,991,026 | MEPS04 |
| Prescription Drug Outlets | | |
| Office visits with Rx in 2004 | 354,919,172 | MEPS04 |
| Stays with Rx at discharge in 2004 | 12,801,928 | MEPS04 |
| Outpatient visits with Rx in 2004 | 15,696,335 | MEPS04 |
| ED visits with Rx in 2004 | 19,008,835 | MEPS04 |
| Population (patients) | | |
| Net increase, mid-2003 to mid-2004 | 2,842,135 | Census |
| Births, mid-2003 to mid-2004 | 4,112,637 | Census |
| Deaths, mid-2003 to mid-2004 | 2,448,288 | Census |
| In-migration, Mar 2003 to Mar 2004 | 1,272,000 | Census |
| Out-migration | 84,214 | Calculated |

SOURCES: MEPS04 – Agency for Healthcare Research and Quality, 2004 Medical Expenditure Panel Survey, 2004; Census – U.S. Census Bureau, various years.

In the estimates below, we assumed that the patient ID is as secure as the health insurance card, driver's license, or credit card that most of us carry. Such cards also offer relatively little security. It is easy for an individual to obtain several cards[2]; worse, it is possible for an individual to use a card that rightfully belongs to someone else (Menn, 2006). But it is expensive to protect against unintentional duplication and unauthorized use of IDs. The Real ID Act is intended to establish such an identifier. The National Governors Association (2006) has estimated that issuing a Real ID—i.e., one ID per person and one person per ID—would cost about $37 per ID. The Governors do not specify how often the ID would need to be renewed.

The Real ID is a token—probably a card—that positively identifies an individual in a face-to-face encounter. It will incorporate a photograph, fingerprint, or other proof of identify. But how can one prove his identity on the Internet? The answer is by employing public-key infrastructure (GAO, 2001). One can purchase a two-year ACES (Access Certificate for Electronic Services) digital certificate for $80 (see, for example, ACES: Access Certificates for Electronic Services homepage, no date), and it serves much the same purpose for electronic transactions as a Real ID would serve for face-to-face transactions.

Clearly, it is costly to guarantee that there is one ID per person and one person per ID. It would cost $11.1 billion over five years to issue a Real ID to each of the 300 million people in the United States and about the same amount per year to keep everyone provided with a current digital certificate. If somebody uses Mrs. Smith's insurance number to obtain health care, and leaves Mrs. Smith with a hefty bill and a corrupted medical record, Mrs. Smith will wish she had spent $40 per year to avoid it. But is it worth $40 per year to you to buy insurance against a one-in-a-million chance of suffering Mrs. Smith's fate?

The system costs we estimate below include only the cost of annually re-enrolling patients in the system and providing them with a card similar to an insurance card: $1.5 billion per year. If the reader wishes to replace this card with a "Real ID," he should add the costs cited above.

## System 1: A National System That Accesses Provider-Held Data

Present-day attempts to share provider-held data occur through regional health information organizations (RHIOs). Examples of RHIOs can be found in Massachusetts (Halamka et al., 2005), Indiana (McDonald et al., 2005), Seattle (Classen et al., 2005), and elsewhere. The Connecting for Health collaborative (2005) has based a design for

---

[2] Hieb (2006) suggests that a patient may want to have multiple IDs, and to associate different information with each of them. One ID might provide access to all PHI. A second might exclude information about care related to mental illness and substance abuse, or to HIV/AIDS care. However, it would not be good for part of a patient's PHI to be unintentionally withheld from a provider.

a national system on the Massachusetts system, and theirs is the design whose cost we will estimate.

In the Massachusetts system, organizations that are members of the RHIO are the sources and hosts of the PHI, and are also the users of the system. The RHIO is the broker. The Massachusetts RHIO has both provider and payer members; we consider only the providers.

### Sources

To participate, a provider organization must maintain its medical records in electronic form. Perhaps 10 to 20 percent of physicians currently have EMRs and, hence, keep their records in electronic form (Connecting for Health, 2005b; Kaushal et al., 2005; National Center for Health Statistics, 2006). Over 80 percent of hospitals have basic IT systems in their radiology, laboratory, and pharmacy departments, but only about 30 percent of hospitals keep records of actual patient care in electronic form. Over the next 10 to 15 years, however, we expect most paper systems to be replaced by electronic ones (Bower, 2005). According to Kaushal et al. (2005), physicians and hospitals will adopt electronic-record systems at a capital cost of $69 billion and an annual operating cost of $17 billion. (Girosi, Meili, and Scoville, 2005) give estimates about half as large.) Skilled nursing facilities, home health agencies, and stand-alone laboratories and pharmacies will incur capital costs for HIT of about $34 billion and annual operating costs of $9 billion. Total system costs to adopt EMR would be a one-time expenditure of $103 billion ($69 billion + $34 billion) and annual cost of $26 billion ($17 billion + $9 billion).

The RHIO will establish standard formats for the information the providers will share; to play the role of source, a provider must conform to those standards. Current EMRs are not interoperable, so providers will need to develop interface software. Kaushal et al. (2005) quotes costs of $15,000 to $90,000 per interface (average: $60,000). Each provider may have several information systems requiring interfaces, so the cost per physician group averages about $130,000 and the cost per hospital, about twice that.

We consider these interface costs to be overestimates, and we (somewhat arbitrarily) cut them in half. At present, a health-information exchange is something for hobbyists, much as the Internet was 10 to 15 years ago. The hobbyists must assemble their own systems as best they can, meaning that only providers who currently have EMRs or purchase them in the near future will have to build or install interfaces. Providers who join the movement late will incur little or no interface cost, because by then vendors will be building the ability to exchange information into their EMRs.

### Hosts

In the Massachusetts organization, providers are expected to host their own data. Hosting one's own data should be much the same as hosting one's own website, which

can be done for a few thousand dollars in one-time expenses and a few hundred dollars per month thereafter. A host will need the hardware on which to store data online (one or more servers with disks, costing perhaps $10,000), an Internet connection with the necessary capacity (e.g., a T1 line with 1.544-megabit-per-sec capacity, for around $300 per month), and what Connecting for Health (2005b) refers to as a clinical data exchange (CDX) gateway. A large provider organization (e.g., an integrated delivery network) might have its own IT department to maintain the service. A small provider (e.g., a small physician practice) might need to outsource the service.

For our estimate of total system cost, we assumed that the one-time cost per provider averages $10,000 and the monthly cost per provider is $400. From Table B.1, we see that there are slightly more than 300,000 providers, or hosts, of all kinds, yielding a total one-time cost of $3 billion, and annual costs thereafter of $1.25 billion.

**Broker**

In Massachusetts, the RHIO is the broker. A user who seeks PHI about a particular patient first submits his query to the RHIO. The RHIO maintains a Record Locator Service (its main function) that identifies which member organizations possess data about that patient and forwards the query to them. Each organization that receives the query must *authenticate* the user (determine whether he is who he claims to be, and what information—if any—he is entitled to receive), and (depending on the outcome) send or display the information. In a national system, at least some of the authentication process may migrate to a national organization. Connecting for Health outlines several variations of this transaction, but this is the simplest.

The Massachusetts RLS maintains a database of pointers to the network location of patient information Each member organization gave the RHIO a database consisting of commonly available data elements (e.g., name, gender, date of birth, zip code, and SSN) describing each of its patients, plus that patient's local medical-record number. The common data elements were compared to determine when medical-record numbers from two member organizations referred to the same patient. Using its own local medical-record number, a member organization can look up other organizations' local medical-record numbers for a patient in the RLS database. Or a provider who is not a RHIO member can provide enough commonly available data elements to uniquely identify the patient, and similarly look up all the local medical-record numbers for that patient.

Such a RLS database is often called a Master Patient Index (MPI). A vendor might ask a one-time payment of a few hundred thousand dollars to link the records of a million patients spread across multiple data sources, and a few million dollars to link the records of many tens of millions of patients.[3] A typical annual maintenance fee is 20 percent of the one-time charge.

---

[3] Personal communication, December 2006.

The Massachusetts system links only a few dozen member organizations. To expand it to a national scope, Connecting for Health (2005a) proposes forming a "network of networks." Individual providers would be linked by local systems, such as the one in Massachusetts. Local systems would be similarly linked by regional systems, each regional system building an MPI from the MPIs of the lower-level systems to which it is linked. Regional systems would be linked by systems of still larger scope. Ultimately these systems would form a multi-level hierarchy encompassed by a single national MPI.

To estimate the cost of all these linking organizations, we must estimate their number. Kaushal et al. (2005) assumed that there would be 307 local RHIOs (Kaushal calls them *central hosts*), five regional RHIOs (*super hosts*), and one national organization (*national host*), with total (for all 313 organizations) one-time costs of about $100 million and annual costs thereafter of $87 million. These costs include everything *except* MPIs for each organization.

We estimate the cost of MPIs for these organizations as follows. Each local RHIO would serve about 1 million patients, so we would estimate a one-time cost of (let us say) $250,000 and an annual maintenance cost of $50,000 to construct an MPI for each local RHIO. The remaining six higher-level RHIOs would need MPIs for tens or hundreds of millions of patients, and since the costs of constructing an MPI depend on its size,[4] we estimate a one-time cost of $2.5 million and an annual maintenance cost of $500,000 for each of them. The total would be a one-time cost of over $90 million and a maintenance cost of over $18 million per year. We assume that Kaushal et al. did not include the MPI costs in their estimates.

Hieb (2006) suggests an alternative method for linking the PHI of a patient in two or more local RHIOs. To form this link, a patient asks his health care provider (it is a voluntary scheme) to issue him a Universal Healthcare Identifier (UHID). The provider requests the UHID from his RHIO, which in turn obtains it from a national organization. This organization maintains a database of all the UHIDs it has issued, together with a pointer to the RHIO to which it was issued. The RHIO keeps track of the patient to whom the UHID belongs.

When the patient seeks care from a provider who is a member of a different RHIO, he presents his UHID. The provider queries the national organization to determine which RHIO to ask for the patient's data, and then queries that RHIO to obtain it. To enable the patient's usual provider to retrieve information from remote providers, the remote providers' RHIO must alert the national organization that the next query on this patient's UHID should return their address as well. Hieb (2006) estimates the total cost of the national organization to be $25 million for the first five years. There will also be a cost for enrolling the patient, which we discuss in the "Users" section that follows.

---

[4] Personal communication, December 2006.

Ideally, one wants to locate 100 percent of a patient's PHI. Both approaches outlined here fall short, because the record-linking algorithms are imperfect. But we should expect the Hieb method to do slightly worse, because it makes no attempt to link historical information to the UHID. (A matching algorithm that uses commonly collected data elements, such as name and date of birth, is the only way we know to link past information.) Such linking will matter only if two or more RHIOs possess PHI for a particular patient before that patient obtains a UHID.

### Users

The users of the Massachusetts system are the people working in or for the RHIO's member organizations. It is the responsibility of the member organizations to determine which of their people have rights to access the system and under what circumstances. But we are considering a national system, and for that we need a systemwide means of validating users. Each user will need to acquire a means to establish his identity to the system, and the ID will generally specify what information the user is allowed to access. For health care professionals, the ID-granting process could be grafted onto the current state licensing process. Annual license fees are generally a few hundred dollars, and, unless we require protections against people posing as doctors that are greatly increased over what present practices provide, we doubt that using licenses in this new way would increase costs significantly.

The cost per transaction (especially per office visit to a physician) is likely to be decisive. At each visit, somebody retrieves the patient's on-site chart and the physician looks at it before seeing the patient. If it requires only a few minutes of extra staff or (especially) physician time to retrieve and examine the off-site information, physicians are unlikely to use the system.[5] One of the barriers to the adoption of electronic records by physicians has been the extra physician time such records seem to demand. Even a couple of extra minutes per patient encounter is unacceptable. To the extent possible, then, the system should be designed so that both on-site information and off-site information are retrieved and displayed in a single, seamless operation.

Patients appear to have no access to their own PHI through the Massachusetts RHIO (Halamka et al., 2005). To use the system, a patient will need to acquire a means of establishing his identity to the system. It would be simplest if the patient's primary care provider (or the provider's office staff) enrolled him, because this provider already knows the patient and can vouch for his identity. The patient's user ID may be something as simple and inexpensive as a username and password, and it would authorize the user to access only information about that patient (and perhaps only some of

---

[5] The cost per transaction is likely to be paid in the currency of provider time, not money. But it can be expressed in equivalent dollars. Patients make about one billion office visits per year, and if this system adds 1 minute of physician time to each visit, that is the equivalent of 7,000 or 8,000 physician-years (depending on the hours per year a physician works). At $200,000 per physician-year, this would be the equivalent of $1.5 billion per year. The cost of extra staff time can be estimated in the same way.

that patient's information). We suppose that enrolling a patient would require 5 minutes of the office staff's time (at $1 per minute) and would be repeated annually.

The Hieb approach would also have to enroll patients. We assumed that enrollment would take the same 5 minutes of staff time. But the patient would also have to be issued a card to present to providers who are members of another RHIO. Considering the length of the UHID (32 digits), it would be wise to put it in a bar code or write it to a magnetic strip on the card. Such cards cost only a few cents to make.

### Total Cost of System 1

We summarize total costs of system 1 in Table B.3.

The only items of expense that are related to patient IDs are the costs of MPIs and the enrollment and transaction costs. Thus, the largest costs have nothing to do with patient identification. (This remains true even if one adds the costs of a Real ID.) User-related costs (enrollment and transactions) are not likely to appear in any accounting system as an item of expense. But they will eat away at physician time and staff time, a few minutes per occasion.

The construction of this system would take years. One-time costs would be spread over those years and would not all come due in a single year. Suppose the system were implemented over ten years. One-tenth of the one-time cost would come due in each year. In year 1, there would be no annual cost. In year 2, one-tenth of the system would be up and running, so one-tenth of the annual cost would have to be paid. The year-3 annual cost would be two-tenths of the total. Year 10 would have the largest cost (about $50 billion). Each year thereafter would have the total annual cost from Table B.3, or $41 billion.

**Table B.3**
**Total Costs of a System to Access Provider Data (System 1)**

| | | Cost ($billions) | |
|---|---|---|---|
| **Role** | **Item of Expense** | **One-Time** | **Annual** |
| Source | Adopt EMRs | 103 | 26 |
| | Interfaces | 26 | 10 |
| Host | Gateway | 3 | 1.44 |
| Broker | RHIOs | 0.1 | 0.087 |
| | MPIs | 0.09 | 0.018 |
| User | Enrollment | | 1.5 |
| | Transactions | | 1.5 |
| Total | | 132.19 | 40.545 |

## System 2: A National System That Accesses Payer-Held Data

Our model for accessing payer-held data (i.e., eligibility, benefits, and claims information) builds on the Blue Cross Blue Shield Association (BCBSA). The BCBSA website (no date) lists 39 member companies operating in all 50 states and the District of Columbia. A member of any BCBSA health plan can obtain services in the region served by any other and have her claims passed on to her home plan. The patient ID is her insurance card, which identifies her plan and her number within the plan—a unique patient identifier. We assess the cost of a system by which any provider could obtain not only a patient's eligibility and benefits information, but also claims information, merely by entering information from the patient's insurance card into a website run by a national organization.[6]

### Sources

The sources of information would be payer organizations (health insurance companies). Payer organizations collect virtually all eligibility, benefit, and claims information in electronic form. Of all claims, 75 percent are submitted electronically, up from 44 percent in 2002 (American Health Insurance Plans [AHIP], Center for Policy and Research, 2006). Of course, claims contain much less information than medical records, so sharing them is less valuable; however, medication lists, some lab values, and the occasional procedure could be useful to the requesting provider. But because claims are already in digital form, a system that shares them can cover a larger fraction of the population in a shorter time than one that shares medical records.

The drawback is that there is a delay from the time a service is performed to the time the claim is posted. Fifty-seven percent of "clean" (complete, properly documented) claims were processed within 7 days, and 98 percent were processed within 30 days (AHIP Center for Policy and Research, 2006). Only the time from the receipt of the claim by the payer is counted; any delay by the provider in submitting the claim is not counted. According to a 2002 AHIP survey (AHIP Center for Policy and Research, 2003), over half of claims were received by payers 15 or more days after the last date of service.

According to Table B.1, the system will have at most a few thousand sources, which agrees well with Kaushal et al.'s estimate of 2000 payers. We use Kaushal's

---

[6] An alternative design for the system could provide access to a patient's data through the same kind of MPI used in the system we discussed above. In 2001, three large pharmacy benefit managers (PBMs)—Advance PCS, Express Scripts, and Medco Health Solutions—jointly founded RxHub, an electronic information-exchange system designed to provide patient-specific eligibility and formulary data and, since 2003, patient-specific medication histories. (Strictly speaking, PBMs are not payers, but they work for payers.) RxHub uses an MPI built from the member PBM's lists of patients, using a matching algorithm. Providers look up data for a patient by entering common data elements to identify the patient. As of 2006, they reported that RxHub had provided data on 160 million patients and had responded to almost 10 million queries about eligibility and 1 million queries about medication histories quarterly (RxHub, 2004 and 2006).

estimates for payer interoperability costs, which equal one-time payer expenditures of $0.37 billion and annual expenditures of $0.15 billion thereafter.

### Hosts

Payers will host their own data. Most insurance companies already have websites, and many already offer providers electronic means of checking the eligibility and benefits of their patients. We assumed costs for offering providers claims data as well would be minimal. The cost of formatting claims data appropriately are included in the interface costs above.

### Broker

The broker would be a trade association, such as BCBSA. Through the association, member health plans would develop any necessary standards—e.g., for claims data formatting or the information to be included on insurance cards. The broker would maintain a database of pointers to the member payers so that it could locate a patient's health plan once a provider entered the patient's insurance information. We anticipate that the provider's query would be automatically forwarded to the appropriate health plan, and the health plan's response would be automatically forwarded to the provider.

We adopted Kaushal et al.'s costs for his super and national hosts for the regional and national organizations that serve as brokers, for a one-time cost of $12 million and annual costs of $7 million thereafter.

Note that, because this system uses each health plan's existing insurance number to identify the patient, the system can only retrieve PHI from a patient's current health plan. If a patient has recently changed health plans, information accumulated by the old plan will not be available. One could make it available by carrying out a record-linking exercise such as the one described for the Massachusetts system.

### Users

The users would be providers, who would need a means of identifying themselves to the system. As we did for system 1, we assumed that the ID-granting process could be grafted onto the current state licensing procedures and that costs would not increase significantly.

Most providers currently take the insurance card as evidence that they will be paid, and they do not engage in any transaction with the payer until they submit the claim. With this system, there could be an extra transaction for each patient encounter. Using the same means as for system 1, we assigned a cost of $1.5 billion per year to these transactions.

With this system, it would be possible to allow patients access to their claims information as well. Many health plans already allow their beneficiaries to open online accounts through which they can access eligibility and benefits information; generally,

however, they provide access only to the financial information relating to claims. It should cost little or nothing to make the clinical information from claims accessible as well.

### Total Cost of System 2

We summarize total costs of system 2 in Table B.4.

This system uses the same mechanism to identify patients and their data that payers currently use: the insurance card and the payer databases that link the card number to the patient's name, address, date of birth, employer, and so forth. Thus, the table shows no new patient ID costs.

As before, transactions are not likely to appear in any accounting system as an item of expense. But they will eat away at physician time and staff time, a few minutes per occasion.

It appears that most of the elements of this system are already in place. The system could be implemented very quickly and at little cost.

## System 3: Personal Health Records

The characteristic that differentiates a personal health record from any other health record is that it is under the control of the patient (Tang et al., 2006). For example, a patient might carry a USB flash drive that holds his critical PHI. Or, a patient might have his PHR stored online and password-protected. When the patient visits his doctor, he can provide access to his PHR.

### Source

At the present time, patients must enter data into their own PHRs. Recently, however, representatives from AHIP and BCBSA announced a Web-based tool that contains a consumer's claims, administrative information, and core health data (Pizzi, 2006). The tool would be hosted by the insurance companies, which would also populate

**Table B.4**
**Total Costs of a System to Access Payer Data (System 2)**

| | | Cost ($billions) | |
|---|---|---|---|
| **Role** | **Item of Expense** | **One-Time** | **Annual** |
| Source | Interoperability | $0.37 | $0.15 |
| Host | | $0 | $0 |
| Broker | Association | $0.012 | $0.007 |
| User | Transactions | | $1.5 |
| Total | | $0.382 | $1.657 |

it with claims and administrative information. Payers could load this information into PHRs, we suppose, at the same cost as sending it to inquiring providers using system 2 (the "Interoperability" line of Table B.4).

Medical-record information must come from providers. Providers would need to adopt EMRs and build interfaces to format information appropriately, just as for system 1. In system 2, however, the content and format would be appropriate for PHRs (see Tang et al., 2006, for suggested content). National costs would be the same as the "Adopt EMRs" and "Interfaces" lines of Table B.3.

Payers and providers could load each item of information (when it became available—a "push" system) into the PHR automatically. Or the patient could be required to request that her PHR be updated (a "pull" system).

Patients could enter symptoms, self-reported behaviors, family histories, and vital signs measured in the home (weight, temperature, peak flow for asthmatics, blood glucose for diabetics).

#### Host

PHRs can be hosted by insurance companies (Tang et al., 2006), employers (Tang et al., 2006; CHCF, 2007b), or independent organizations.

#### Broker

The patient is the broker. Only if the patient gives a health care provider her PHR (if it is on a USB drive) or the password (if it is online) can the provider access it. In this system, the broker's function costs nothing.

Since the patient controls the PHR, she can accumulate her PHI in it regardless of whether she changes doctors or health plans.

#### User

Patients and providers are the users. Implementing PHRs does not require a new patient-identification function. Updating a PHR will simply be another transaction between one payer or provider and one patient, and payers and providers already have mechanisms for identifying patients.

MedicAlert (2008) memberships cost $40 per year and include a PHR. (MedicAlert asserts that the PHR is at the heart of every membership.) For the entire population of the United States, this would amount to $12 billion per year, although many people might elect not to have a membership. A membership includes more than a PHR, of course. MedicAlert will relay a PHI to emergency personnel as needed. Or it can be stored on an e-HealthKEY, a USB flash drive that stores your PHR data plus software for launching the information automatically once the key is inserted into a USB socket on any computer. MedicAlert will sell a member the e-HealthKEY for $50. We assumed that these costs are typical for a PHR.

There will be a transaction cost each time a provider accesses a PHR, however. Using the same means as for system 1, we estimated this cost at $1.5 billion per year.

### Total Cost of System 3

We summarize total costs of system 3 in Table B.5.

The table shows no new patient ID costs, though one might interpret the username and password that protect an online PHR to be a patient ID. Payers and providers identify the patient in the same way they do now, and sending information to a patient's PHR is just another kind of transaction.

As before, transactions are not likely to appear in any accounting system as an item of expense. But they will eat away at physician time and staff time, a few minutes per instance.

**Table B.5**
**Total Costs of a System of Personal Health Records (System 3)**

| | | Cost ($billions) | |
|---|---|---|---|
| **Role** | **Item of Expense** | **One-Time** | **Annual** |
| Source | Adopt EMRs | $103 | $26 |
| | Interoperability | $0.37 | $0.15 |
| | Interfaces | $26 | $10 |
| Host | | $0 | $0 |
| Broker | | $0 | $0 |
| User | PHR fee | | $12 |
| | Transactions | | $1.5 |
| Total | | $129.37 | $49.65 |

# Bibliography


ACES: Access Certificates for Electronic Services homepage, no date. As of September 1, 2008:
http://www.aces.orc.com/

ACLU—*see* American Civil Liberties Union.

ACS EDI Gateway, Inc., *ANSI ASC X12N 270/271 (Version 4010A) Benefit Eligibility Inquiry/Response Transactions: Washington State Medical Assistance Administration Companion Guide*, Tallahassee, Fla., October 13, 2005.

Agency for Healthcare Research and Quality, 2004 Medical Expenditure Panel Survey (MEPS), 2004. As of September 8, 2008:
www.meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp

AHA—*see* American Hospital Association.

AHIC—*see* American Health Information Community.

AHIP—*see* American Health Insurance Plans.

Alliance, "The National Alliance for Health Information Technology Calls for Creation of Voluntary Unique Patient Identifiers for Exchanging Electronic Health Records," December 13, 2007. As of January 10, 2008, member-only website:
http://www.nahit.org/cms/index.php?option=com_content&task=view&id=328&Itemid=214

American Civil Liberties Union (ACLU), "ACLU Tells CDC New Guidelines for HIV Tracking Violate Privacy, Ignore Public Health Research," press release, New York, January 12, 1999. As of March 25, 2007:
http://www.aclu.org/privacy/medical/15084prs19990112.html

American Health Information Community (AHIC), American Health Information Community: Breakthroughs, "2006 Breakthrough Focus Areas," 2006. As of April 2, 2007:
http://www.hhs.gov/healthit/community/breakthroughs/

American Health Insurance Plans (AHIP), Center for Policy and Research, *An Updated Survey of Health Care Claims Receipt and Processing Times*, Washington, D.C., 2006. As of December 23, 2006:
http://www.ahipresearch.org

———, *Results from an HIAA Survey on Claims Payment Processes*, Washington, D.C., 2003. As of January 9, 2007:
http://www.ahipresearch.org

American Hospital Association (AHA), "Protecting and Improving Care for Patients and Communities: Health Information Technology," 2006 Advocacy Position Paper, Washington, D.C., 2006. As of March 20, 2007:
http://www.aha.org/aha/content/2006/pdf/Iss-Paper-Health-IT-06.pdf

American Psychoanalytic Association, "Position Statement–Confidentiality," last revised September 7, 1999. As of March 30, 2007:
http://www.apsa.org/ABOUTAPSAA/POSITIONSTATEMENTS/CONFIDENTIALITY/tabid/474/Default.aspx

American Standards for Testing and Materials (ASTM), *Standard Guide for Properties of a Universal Healthcare Identifier* (UHID), West Conshohocken, Pa.: ASTM, E1714-00, October 10, 2000. As of March 16, 2007:
http://www.astm.org/cgi-bin/SoftCart.exe/STORE/filtrexx40.cgi?U+mystore+ubkf5409+-L+FUNCTIONS::OF::A::PATIENT::IDENTIFIER:+/usr6/htdocs/astm.org/DATABASE.CART/REDLINE_PAGES/E1714.htm

———, *Standard Guide for Confidentiality, Privacy, Access, and Data Security: Principles for Health Information Including Electronic Health Records*, West Conshohocken, Pa.: ASTM International, 2004.

———, Subcommittee E31.35, *ASTM WK11238—New Guide for the Implementation of a Voluntary Universal Healthcare Identification System*, April 28, 2006. As of January 10, 2008:
http://www.astm.org/cgi-bin/SoftCart.exe/DATABASE.CART/WORKITEMS/WK11238.htm?L+mystore+ecjq9139

Appavu, S. I., *Analysis of Unique Patient Identifier Options: Final Report*, prepared for the U.S. Department of Health and Human Services, November 24, 1997a. As of March 20, 2007:
http://www.ncvhs.hhs.gov/app0.htm

———, quoting from the National Research Council, *For the Record: Protecting Electronic Health Information*, Washington, D.C.: National Academy of Sciences, 1997b.

ASTM—*see* American Standards for Testing and Materials.

Bigelow, James, Kateryna Fonkych, Constance Fung, and Jason Wang, *Analysis of Healthcare Interventions That Change Patient Trajectories*, Santa Monica, Calif.: RAND Corporation, MG-408-HLTH, 2005. As of October 3, 2007:
http://www.rand.org/pubs/monographs/MG408

Blue Cross Blue Shield Association website, no date. As of September 1, 2008:
http://www.bcbs.com/coverage/find/plan

Bovbjerg, Barbara, *Federal and State Laws Restrict Use of SSNs, Yet Gaps Remain*, Government Accountability Office (GAO) Testimony Before the Committee on Consumer Affairs and Protection and Committee on Governmental Operations, New York Assembly, Washington, D.C.: GAO, GAO-05-1016T, September 15, 2006.

Bower, Anthony, *The Diffusion and Value of Healthcare Information Technology*, Santa Monica, Calif.: RAND Corporation, MG-272-1-HLTH, 2005. As of September 1, 2008:
http://www.rand.org/pubs/monographs/MG272-1

Cain, Mary M., Jane Sarasohn-Kahn, and Jennifer C. Wayne, *Health e-People: The Online Consumer Experience*, Menlo Park, Calif.: California HealthCare Foundation, August 2000. As of October 3, 2007:
http://www.chcf.org/documents/ihealth/HealthEPeople.pdf

California HealthCare Foundation (CHCF), *Santa Barbara County Care Data Exchange*, Menlo Park, Calif., April 2007a. As of October 3, 2007:
http://www.chcf.org/topics/view.cfm?itemID=132846

———, National Consumer Health Privacy Survey 2005, Menlo Park, Calif., November 2005. As of October 10, 2007:
http://www.chcf.org/documents/ihealth/ConsumerPrivacy2005Slides.pdf

———, Personal Health Records: Employers Proceed with Caution, Menlo Park, Calif.: Issue Brief, 2007b. As of September 1, 2008:
http://www.chcf.org/topics/view.cfm?itemID=129920

CAQH—*see* Council for Affordable Quality Healthcare.

CCHIT—*see* Certification Commission for Healthcare Information Technology.

Certification Commission for Healthcare Information Technology (CCHIT) homepage, July 2007. As of October 30, 2007:
http://www.cchit.org

CHCF—*see* California HealthCare Foundation.

Classen, D. C., M. Kanbouwa, D. Will, J. Casper, J. Lewin, and J. Walker, "The Patient Safety Institute Demonstration Project: A Model for Implementing a Local Health Information Infrastructure," *Journal of Healthcare Information Management*, Vol. 19, No. 4, 2005, pp. 75–86. As of April 18, 2007:
http://www.ptsafety.org/NewsAndArticles/PSI_DemoProject.pdf

Coalition for Patient Privacy, April 5 2006 Letter to Congress. As of September 1, 2008:
http://usacm.acm.org/usacm/weblog/wp-content/PPR_group_letter.pdf

———, 2007 Patient Privacy Principles. As of August 28, 2007:
http://www.patientprivacyrights.org/site/PageServer?pagename=PrivacyCoalition

———, "47 State & National Organizations, Health IT Companies Join Forces to Demand Consumers Regain Control of Their Personal Health Records," press release, October 18, 2007. As of January 10, 2008:
http://www.patientprivacyrights.org/site/DocServer/CoalitionLetter_Press_Release.pdf?docID=2282

Connecting for Health, *Record Locator Service: Technical Background from the Massachusetts Prototype Community*, New York: Markle Foundation, 2005a. As of February 18, 2007:
http://www.connectingforhealth.org/commonframework/#guide

———, website, Technical Guides: How Information Is Exchanged, no date. As of March 18, 2007:
http://www.connectingforhealth.org/commonframework/technical.html

———, Working Group for Accurately Linking Information for Health Care Quality and Safety, *Linking Health Care Information: Proposed Methods for Improving Care and Protecting Privacy*, New York: Markle Foundation, February 2005b. As of August 15, 2005:
http://www.connectingforhealth.org/assets/reports/linking_report_2_2005.pdf

Council for Affordable Quality Healthcare (CAQH) homepage, July 2007. As of August 31, 2008:
http://www.caqh.org/ucd.php

E-Health Initiative, *Blueprint: Building Consensus for Common Action, Phase I*, "Managing Privacy, Security and Confidentiality," Washington, D.C., October 10, 2007, pp. 75–82. As of October 15, 2007:
http://www.ehealthinitiative.org/blueprint/eHiBlueprint-BuildingConsensusForCommonAction.pdf

Erikson, Jane, "Arizona Law Forbids Use of Social Security Numbers for Identification," *The Arizona Daily Star*, December 29, 2004.

Ferris, N., "Survey Shows Public Distrusts HIPAA; Researchers Detest It," *Government Health IT*, October 2, 2007. As of October 15, 2007:
http://www.govhealthit.com/online/news/350058-1.html

GAO—*see* General Accounting Office for documents published before 2004; see Government Accountability Office for documents published after 2004.

General Accounting Office (GAO), *Information Security: Advances and Remaining Challenges to Adoption of Public Key Infrastructure*, GAO-01-277, 2001. As of April 21, 2007:
http://www.gao.gov/new.items/d01277.pdf

Girosi, Federico, Robin Meili, and Richard Scoville, *Extrapolating Evidence of Health Information Technology Savings and Costs*, Santa Monica, Calif.: RAND Corporation, MG-410-HLTH, 2005. As of October 3, 2007:
http://www.rand.org/pubs/monographs/MG410

Government Accountability Office (GAO), *Health Information Technology: Early Efforts Initiated But Comprehensive Privacy Approach Needed for National Strategy*, Washington, D.C., GAO-07-400T, February 1, 2007. As of March 31, 2007:
http://www.gao.gov/new.items/d07400t.pdf

Grannis, Shaun J., J. Marc Overhage, and Clement McDonald, "Real World Performance of Approximate String Comparators for Use in Patient Matching," *MEDINFO 2004*, Amsterdam: IOS Press, 2004, pp. 43–47.

Greenberg, Michael D., and Susan Ridgely, "Patient Identifiers and the National Health Information Network: Debunking a False Front in the Privacy Wars," *Journal of Health and Biomedical Law*, Vol. 4, No. 1, 2008, pp. 31–68.

Halamka, J., M. Atranow, C. Ascenzo, D. Bates, G. Debor, J. Glaser, A. Goroll, J. Stowe, M. Tripath, and G. Vineyard, "Health Care IT Collaboration in Massachusetts: The Experience of Creating Regional Connectivity," *Journal of the American Medical Informatics Association*, Vol. 12, No. 6, 2005, pp. 596–601.

Harris Interactive, Harris Interactive Poll #27, March 6, 2007b. As of April 3, 2007:
http://www.harrisinteractive.com/harris_poll/printerfriend/index.asp?PID=743

———, How the Public Views Privacy and Health Research, survey for the Institute of Medicine, 2007a; revised and expanded March 2008. As of September 2, 2008:
http://www.patientprivacyrights.org/site/PageServer?pagename=Polls#IOM

Health Information and Management System Society (HIMSS), "*Voluntary Patient Identifier Resolution Position Statement,"* December 18, 2003. *NOTE: This HIMSS resolution is not accessible on the HIMSS website, but it is repeated in the press release applauding an IOM position on the subject, as of January 8, 2008: http://www.himss.org/advocacy/ContentRedirector.asp?ContentId=40023*

Health Insurance Portability and Accountability Act of 1996 (HIPAA), Public Law 104-191, *U.S. Statutes at Large* 110, 1996, §1936.

Health Level Seven website, Clinical Context Object Workgroup (CCOW), "Goals and Standards," July 2006. As of October 31, 2007:
http://www.hl7.org/special/Committees/ccow_sigvi.htm

Healthcare Information Technology Standards Panel (HITSP), *Manage Sharing of Documents Transaction Package*, Version 2, Table 3.1-1, "List of Standards," LIS09A, 2001, as well as Integrating the Healthcare Enterprise (IHE)'s *IT Infrastructure Technical Framework*, Volumes 1 and 2, August 22, 2007. As of September 8, 2008:
www.webstore.ansi.org

HHS—*see* U.S. Department of Health and Human Services.

Hieb, Barry, "The Case for a Voluntary National Healthcare Identifier," *Journal of ASTM International*, Vol. 3, No. 2, 2006.

Hillestad, Richard, James Bigelow, Anthony Bower, Federico Girosi, Robin Meili, Richard Scoville, and Roger Taylor, "Can Electronic Medical Record Systems Transform Health Care? Potential Health Benefits, Savings, and Costs," *Health Affairs*, Vol. 24, No. 5, September/October 2005, pp. 1103–1117.

HIMSS—*see* Health Information and Management System Society.

HIPAA—*see* Health Insurance Portability and Accountability Act of 1996.

HITSP—*see* Healthcare Information Technology Standards Panel.

HL7—*see* Health Level Seven.

IHE—*see* Integrating the Healthcare Enterprise.

Initiate Systems, *Integrating Patient Medical Records in Pursuit of the EMR*, WPEMR-1207, 2008. As of September 2, 2008:
http://www.initiatesystems.com/resources/Pages/default.aspx

Institute of Medicine (IOM), Committee on Data Standards for Patient Safety, *Patient Safety: Achieving a New Standard of Care*, P. Aspden, et al., eds., Washington, D.C.: National Academies Press, 2003.

———, Committee on the Disposition of the Air Force Health Study, *Disposition of the Air Force Health Study,* Washington, D.C.: National Academies Press, 2006.

Integrating the Healthcare Enterprise (IHE), *IT Infrastructure Technical Framework*, Volume 1, *Integration Profiles*, August 22, 2007. As of August 31, 2008:
http://www.ihe.net/technical_framework/upload/IHE_ITI_TF_4_0_Vol1_FT_2007_08_22.pdf

———, *IT Infrastructure Technical Framework*, Volume 2, *Transactions*, August 22, 2007. As of August 31, 2008:
http://www.ihe.net/Technical_Framework/upload/IHE_ITI_TF_4.0_Vol2_FT_2007-08-22.pdf

Integrating the Healthcare Enterprise website, Technical Frameworks, various dates. As of March 18, 2007:
http://www.ihe.net/technical_framework/index.cfm#it

IOM—*see* Institute of Medicine.

Kaushal, R., D. Blumenthal, E. G. Poon, A. K. Jha, C. Franz, B. Middleton, J. Glaser, G. J. Kuperman, M. Christino, R. Fernandopulle, J. P. Newhouse, and D. W. Bates, "The Costs of a National Health Information Network," *Annals of Internal Medicine*, Vol. 143, No. 3, 2005, pp. 165–173.

Lake Research Partners (LRP) and American Viewpoint, National Survey on Electronic Personal Health Records, November 2006. NOTE: This survey was done in preparation for the Markle Foundation's conference, *Connecting Americans to Their Health Care: Empowered Consumers, Personal Health Records and Emerging Technologies*, December 7–8, 2006, Washington, D.C.

Lau, P. Behrendt, M. Bruun-Rasmussen, and K. Bernstein, *Study on Patient Identity in eHealth*, Final Report, Brussels, Belgium: European Commission–DG Information Society & Media, December 21, 2006.

Markle Foundation, National Survey on Electronic Personal Health Records, Washington, D.C., December 2006. As of March 24, 2007, a summary was available at:
http://www.markle.org/downloadable_assets/research_doc_120706.pdf

McDonald, C. J., J. M. Overhage, M. Barnes, G. Schadow, L. Blevins, P. R. Dexter, and B. Mamlin, "The Indiana Network for Patient Care: A Working Local Health Information Infrastructure," *Health Affairs*, Vol. 24, No. 4, 2005, pp. 1214–1220. As of April 18, 2007:
http://content.healthaffairs.org/content/vol24/issue5

MedicAlert website, 2008. As of September 1, 2008:
http://www.medicalert.org

Menn, J., "ID Theft Impacts Medical Records; Victims Face Bogus Bills and Risk Injury or Death. Privacy Laws Make Such Fraud Hard to Pursue," *Los Angeles Times*, September 25, 2006.

Milbourn, Mary Ann, "California to Clamp Down on Social Security Number Use by Business," *Orange County Register*, June 11, 2002.

"More UCLA Medical Center Employees Peeked at Celebrities' Records, State Says," *Los Angeles Times*, August 5, 2008.

National Center for Health Statistics (NCHS), *Health United States, 2006*, Hyattsville, Md., 2006. As of April 13, 2007:
http://www.cdc.gov/nchs/hus.htm

National Committee on Vital and Health Statistics (NCVHS), Subcommittee on Standards and Security, *Hearing Minutes*, July 20–21, 1998, Chicago, Ill. As of March 5, 2007:
http://ncvhs.hhs.gov/980720mn.htm

National Governors Association, *The Real ID Act: National Impact Analysis*, September 2006. As of January 22, 2007:
http://www.nga.org/Files/pdf/0609REALID.PDF

NCHS—*see* National Center for Health Statistics.

NCVHS—*see* National Committee on Vital and Health Statistics.

Patient Privacy Rights website, "47 State & National Organizations, Health IT Companies Join Forces to Demand Consumers Regain Control of Their Personal Health Records," October 18, 2007. As of January 10, 2008:
http://www.patientprivacyrights.org/site/DocServer/CoalitionLetter_Press_Release.pdf?docID=2282

Piasane, Susan, Vice President of Communications, AHIP, personal communication, August 19, 2008.

Pizzi, R., "Insurance Industry Reps Reveal PHR Plan," *Healthcare IT News*, December 13, 2006. As of April 16, 2007:
http://www.healthcareitnews.com

Row, G. Leg, and J. Metzger, *E-Disease Management*, Falls Church, Va.: First Consulting Group, November 2001. As of January 10, 2008:
http://www.chcf.org/documents/ihealth/EDiseaseManagement.pdf

RxHub, *The Opportunity and Challenge of RxHubMEDS*. Part I: *Background and Benefits*, 2004. As of April 13, 2007:
http://www.rxhub.net

———, *The ROI Behind ePrescribing: Cost Savings, Patient Safety and Physician Adoption*. RxHub Symposium Summary, May 9–10, 2006. As of April 13, 2007:
http://www.rxhub.net

Social Security Administration (SSA), Social Security Death Master File (SSDMF), updated quarterly. As of October 3, 2007:
http://www.ssdmf.com

———, "Enhanced Social Security Card Prototype," *Report to Congress on Options for Enhancing the Social Security Card*, Washington, D.C., 1997, Chapter 5. As of October 10, 2007:
http://www.ssa.gov/history/reports/ssnreportc5.html

SSA—*see* Social Security Administration.

Stanley, J., and B. Steinhardt, *Bigger Monster, Weaker Chains: The Growth of an American Surveillance Society*, New York: ACLU, Technology and Liberty Program, January 15, 2003. As of March 26, 2007:
http://www.aclu.org/privacy/gen/15162pub20030115.html
Also, as of August 28, 2008:
http://www.aclu.org/FilesPDFs/aclu_report_bigger_monster_weaker_chains.pdf

Stubbs, J., American College of Physicians, "Comments on the Medicare Program: E-Prescribing and the Prescription Drug Program Proposed Rule (42 CFR 423)" letter to Secretary, HHS, April 1, 2005. As of March 30, 2007:
http://www.acponline.org/hpp/eprescribe_com.pdf

Tang, P. C., J. S. Ash, D. W. Bates, J. M. Overhage, and D. Z. Sands, "Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption," *Journal of the American Medical Informatics Association*, Vol. 13, No. 2, 2006, pp. 121–126.

U.S. Census Bureau, homepage, no date. As of September 8, 2008:
http://www.census.gov

U.S. Department of Health and Human Services (HHS), *Unique Health Identifier for Individuals: A White Paper*, Washington, D.C., 1998. As of June 19, 2007:
http://www.epic.org/privacy/medical/hhs-id-798.html

———, Office of the National Coordinator homepage, no date. As of June 25, 2007:
http://www.hhs.gov/healthit/onc/mission

Von Schelling, Herman, "Coupon Collecting for Unequal Probabilities," *The American Mathematical Monthly*, Vol. 61, No. 5. May 1954, pp. 306–311.