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

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.