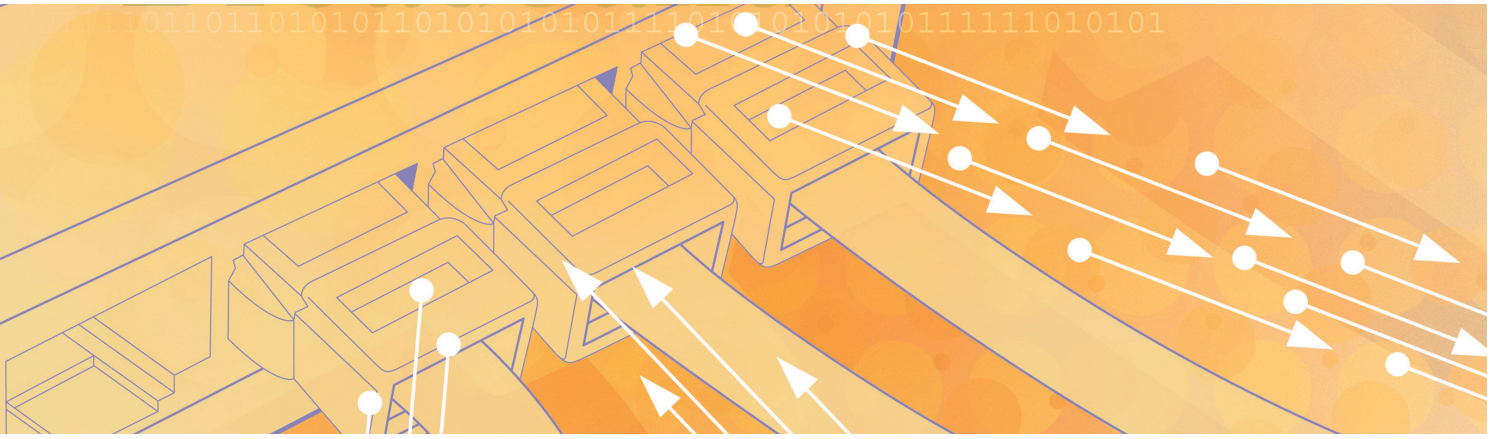
The Future of Patient Identification



Once data can move across vast networks, how to find it? The accuracy and flexibility of patient matching will determine interoperability's success.

by **Lorraine Fernandes,** RHIA, and **Michele O'Connor,** MPA, RHIA

Christina Jones, a Chicago resident, is spending the day with a friend in nearby Milwaukee. While shopping after lunch, she collapses and is rushed to the local emergency department. Since Christina is unconscious, the doctor can't ask her if she's currently taking or is allergic to any medicines, and her friend doesn't know with any certainty.

Fortunately, the hospital has invested in technology that enables the exchange of patient data between healthcare providers. The registration personnel are able to determine that Christina has records at a Chicago facility and can access critical allergy and drug interaction information, helping the emergency staff provide the most complete care. Christina recovers from the stroke, but without access to her pertinent medical history, the outcome could have been quite different.

Key to retrieving Christina's medication history was the

ability of the registration staff to identify her record in another provider's system. Patient identification is fundamental to the interoperability of health data and the resulting promise of improved care. Although an exact match on a name or a unique identification number may seem the surest method for identifying a patient's dispersed records, sophisticated searches that match a set of demographic data may prove better adapted to the current climate and complexity.

Matching: Key to Interoperability

From an HIM point of view, linking patients to their scattered records is currently a tedious, often inaccurate, and very slow process, because data are housed in systems that often cannot talk to each other—the result of siloed, proprietary legacy systems. In many cases, the patient relating her or his own medical history becomes the primary means for retrieving historical information. Interoperability initiatives seek to solve this problem by fostering cooperation between systems and facilities and enabling them to work together for improved patient care, safety, and decreased healthcare costs. Health information managers will both greatly influence and be affected by the rise of interoperability, as it transforms how records are stored, indexed, and distributed.

Patient identification and matching hold one of the most important keys to achieving true interoperability. To achieve the highest level of accurate patient identification, patient matching must determine whether or not records are systematically linked across disparate systems. As a key component for electronic health records (EHRs), patient matching links records to provide the most complete, accurate picture of a patient's medical history. "To achieve high-fidelity matches, ideally identifiers create an accurate and unique representation of a patient, regardless of population size," explains Shaun Grannis, MD, MS, a research scientist with the Regenstrief Institute at the Indiana University School of Medicine. The risks of failing to match patient records—or doing so incompletely or inaccurately—are many, including poor patient care and satisfaction, increased costs, regulatory noncompliance, and potential legal liability.

The Office of the National Coordinator for Health Information Technology, established in part to meet President Bush's goal of widespread interoperable EHR adoption, views interoperability as the key to a nationwide network of EHRs and information exchange, achievable by harmonizing standards and harnessing available Internet infrastructure. Based on the results of its January 2005 request for information, the office concluded that patient matching is a primary interoperability issue. A survey by eHealth Initiative found similar results, with accurate patient matching and identification in data exchange as the second most pressing issue behind financial sustainability.[1] Connecting for Health, a public-private collaborative established by the Markle Foundation, has also concluded that accurately matching records enables the searching and sharing that are fundamental to interoperability.[2]

Unique Problems with Unique Identifiers

All integrated delivery networks (IDNs), regional health information organizations (RHIOs), and other data exchange organizations must address the challenges of accuracy and patient matching early in their formation, before they can embark on successful data exchange. Regardless of whether a group relies on a central data repository or a federated or hybrid model, its members must establish a strategy for accurate patient matching.

Patient matching techniques have evolved along with processor speed and database technologies. Historically, healthcare organizations employed a deterministic matching approach that looked at a few characters in key fields. However, the method fails to scale, or adapt, well as the volume of data increases.

Matching records via a unique identifier such as a patient's Social Security number (SSN) or a new national number is not a simple solution to the coordination problem. Adding new data to a record does not solve the problem of inconsistent or incorrect information that may already be there. Numerous published studies find average error rates of 5 to 10 percent even in single-facility MPIs.

"A far better use of the energy that would go into collecting new data should instead go to cleaning up existing data, hardest cases first," says Clay Shirky, chair of a Connecting for Health technical subcommittee and an adjunct professor at New York University, who has directed extensive research on the healthcare technology issues surrounding interoperability. Doing so will greatly improve match rates without need for additional data.

Concerns over Privacy, Usefulness, Logistics

Some systems allow the SSN to be used for search purposes but limit its display to the last four digits or block it entirely. However, this use conflicts with many privacy concerns and some new legislation, and it could potentially entice hackers to try to steal patient identities.

When SSNs were introduced in the 1930s, their use was limited to pension and retirement purposes, with explicit instructions against use for identification. Today, use and abuse of SSNs are rampant. A new healthcare identifier may begin in a similar way—explicitly for healthcare purposes—but it may not take long before its purpose is corrupted.

There are additional concerns beyond privacy. "The first big problem with the SSN is that it has no checksum, so there is no way to tell if a number is wrong without reference to the original," says Shirky. "This problem has no solution. The second problem is that it is both an identifier and an authenticator. My SSN points to me, but to prove I'm me, I have to disclose my number. Thus use of the SSN degrades its value." Shirky further notes that unique identifiers can fall victim to poor data entry just the same as other identifiers, such as name or date of birth.

Developing and disseminating a new national healthcare identifier would be extremely costly, and it would take years. Many legacy systems that have millions of existing medical records would not be able to accommodate a new mandatory field. The US government is unlikely to fund a "rip and replace" strategy

to implement a mandatory identifier, and without full adoption, interoperability could not be achieved. A unique identifier is not the panacea to patient identification and interoperability. Interoperability is possible today without a unique identifier by using readily available patient-matching technology that employs probabilistic algorithms.

Meeting the Challenges of Scale and Complexity

Adding new data elements, cleaning an existing MPI, or creating and disseminating a unique healthcare identifier will not single-handedly lead to interoperability. Rather, effective interoperability rests on the combination of clean data and the ability to accurately and quickly match and link patient information across records.

Managing patient identities across the diverse healthcare system requires a flexible, open, interoperable architecture that is scalable. Interoperability should not require healthcare organizations to rip out and replace legacy systems that may have inconsistent, incomplete, or fragmented data. Instead, interoperability should enable these legacy systems to share data, despite differences that exist across systems or even within a single vendor's architecture. The architecture must also be adaptable to both current standards and needs (which are sometimes contradictory) and their inevitable evolution.

Probabilistic Matching

As healthcare delivery systems face increasing data volumes with multiple matching attributes, many organizations have adopted probabilistic matching algorithms that offer a more accurate, dynamic, and robust matching approach. Rather than relying on an exact match on a name or a unique identifier, probabilistic matching relies on a combination of readily available data, such as name, birth date, zip code, and address.

Probabilistic matching can improve the rate and quality of matched records by considering a database's specific, unique characteristics.[3] The more fields it compares, the greater its ability to determine accurate and false matches.[4] Research recommends probabilistic matching for situations requiring greater sensitivity, or overall accuracy.[5]

Patient-matching software using probabilistic algorithms can be intelligent enough to overcome common data entry errors, such as misspellings, transposed digits, or "thin" data stemming from limited data capture. It can also take into account variances such as nicknames, so that records for *Christina Jones* and *Tina Jones* are linked if other demographic information matches. Other robust techniques are required to manage hyphenated names, as well as non-Western naming conventions.

The probabilistic logic in patient-matching technology drives algorithms to weigh frequency and uniqueness of data, enabling it to be tuned to find the best possible results based on the available data. As Grannis notes, "There are opportunities to leverage name frequencies in matching; for instance, a last name of *Smith* conveys less assurance of a match than *Zielwicki*." Regional differences can also be taken into account, so a match on the name *Jorge Rodriguez* would score very differently in South Dakota than in Southern California or New York City. With probabilistic logic and robust database architecture, there should be no trade off between accuracy and scalability—the logic is extraordinarily capable of handling growing files and databases with no sacrifice in accuracy, results, or performance.

For example, when matching on names, probabilistic algorithms test for all possible name alignments and go beyond exact matches to consider nicknames, phonetics (e.g., *Christina* versus *Kristina*), transposed last and first names, or the use of initials (e.g., *Christina L. Jones* versus *Christina Louise Jones* or *C. Louise Jones*). By using observed frequencies that address commonly occurring attributes, probabilistic algorithms examine and use the available data. This is essential, since healthcare organizations differ greatly in how they capture and store data. After extensive research, Connecting for Health recommended probabilistic algorithms for patient matching, citing numerous examples of their success in large-scale installations.[6]

RxHub is one organization employing probabilistic matching. The company collaborates with the nation's three largest prescription benefit managers, facilitating the exchange of prescription and benefit information for more than 150 million individuals. By employing software that uses probabilistic logic, RxHub's network of prescription benefit managers, physicians, and pharmacies can match patient records across numerous, ever-changing sources. As a result, clinicians have the most up-to-date information possible about a patient's coverage and medication history, with subsecond response times when searching across a database of hundreds of millions of records.[7]

However, not all probabilistic algorithms are created equal, nor will the same algorithm applied to different situations yield identical results. Most are tunable to achieve different specific false positive (missed linkage) and false negative (incorrect linkage) rates. Hence, HIM professionals must carefully evaluate the accuracy, scalability, performance, and current deployments of probabilistic algorithms as they evaluate the patient identification component of interoperability. HIM professionals possess a level of experience and attention to detail that will greatly enhance the success of an interoperable model that focuses on accurate identification while simultaneously affording the highest level of data privacy possible.

Quality Data Attributes Needed

Without a solid foundation of accurate patient identification, a clinician using an EHR will have incomplete and inaccurate results. The worst case scenario could become a real possibility, with clinicians relying on inaccurate identification to provide care. Data quality and completeness have a large impact on false-negative rates.

Consider an IDN database of more than three million records with four attributes widely available for matching: name, date of birth, zip code, and SSN. The table "Better Data, Better Matches" on page 40 shows sample scenarios in which the four attributes are present in different percentages. Looking at the data, the best (lowest) false-negative rate occurs in scenario D, when 100 percent of names and 90 percent of dates of birth, zip codes, and SSNs are present. The worst (highest) false-negative rate appeared in scenario B, when 100 percent of names and 90 percent of both dates of birth and zip codes were present, but SSNs were lacking.

Better Data, Better Matches

Scenario	Percent of Attributes Present				False-negative Rate (%)
	Name	Date of Birth	Zip Code	SSN (last 4 digits)	
A	100	100	100	0	6
B	100	90	90	0	22
C	100	90	90	70	7
D	100	90	90	90	3

Differing scenarios illustrate that the greater the number and completeness of data elements present for matching patient records, the lower the error rate (scenario D). When fewer elements are present, the rate of bad matches increases (scenario B).

Source: Initiate Systems, Chicago, IL

In this example, adding the SSN provides a fourth point of matching and improves the results. Using just the last four digits of a SSN (as some organizations have already opted and others are considering, especially as legislation puts new restrictions on use of the SSN) provides most of the patient identification benefit available with the full SSN. While more "collisions" will occur (since the likelihood of people sharing the last four digits of an SSN is higher than for the full nine-digit number), the overall impact is small. According to Scott Schumacher, PhD, a data scientist at Initiate Systems, the false-negative rate would only rise one percentage point, to 8 percent in scenario C.

More important than any single piece of information is the overall breadth and content of data. "If [the] SSN is not present, we require much tighter agreement on remaining identifiers…. The more consistently and accurately recorded identifiers are available for each patient, the more accurate the linkages," Grannis notes.

Facilities should agree on using certain fields for the purposes of matching. While consensus is high for using some fields—name, date of birth, zip code—other data elements are not uniformly collected. What should be collected? Mother's maiden name? Telephone number? Place of birth? Aliases?

"Almost anything can be used," Schumacher explains. "But it needs to be something widely collected by hospitals, pharmacies, physician offices, and other points of care." Data elements widely used in probabilistic matching include name, aliases or nicknames, date of birth, zip code, mother's maiden name, SSN, and telephone number. Possible additional elements include retinal scans and thumbprints.

The problem lies in coordinating facilities and networks to ensure that the data elements are useful and populated in enough places to make them valuable attributes. When an attribute is not commonly collected, its value is lessened. Grannis cites an example: "A recent review of two large healthcare organizations revealed that SSN is present for 60 to 65 percent of all patients. This percentage drops off precipitously in children. Consequently, pediatric patient matching is even more challenging." Shirky elaborates, saying, "If Dr. Schumacher starts collecting zip+4, I start collecting mother's maiden name, and Dr. Grannis starts doing retinal scans, then we've made our individual databases better but have done nothing for the coordination problem."

Several federal agencies and public-private collaboratives are leading the discussion about which data elements most effectively enhance data quality and matching. The Commission on Systemic Interoperability identified the need for a national standard on patient identification as a priority in reaching interoperability, and it urged the industry to resolve the debate and choose a method [see story page 26].

Connecting for Health's Data Standards Working Group is one group studying the potentials of a minimum data set to ensure interoperability between regional systems and other health organizations. Such a standard might require that name (first and last), date of birth, gender, and zip code be used for matching; supplemental fields used by some algorithms might also be considered, such as phone number (which is far richer today than 10 years ago), address, or SSN, either complete or partial. The question driving the discussion is which fields or combination of fields best identify a person as unique.

Ultimately it is not just the data fields that are important, but the matching method and the speed of searching and linking the available data. Successfully addressing all three points will enable healthcare organizations to create interoperable EHRs. HIM professionals involved at all levels of RHIO and IDN development play a crucial role in ensuring that plans for EHRs and interoperability are achievable and sustainable. Interoperability—and the improved patient care that comes with it—will succeed with the collaborative efforts of HIM professionals dedicated to the highest levels of patient safety and care. ❖

Notes

1. eHealth Initiative. "eHI Foundation Releases National Report on Health Information Exchange (HIE) Efforts." Press release. August 29, 2005. Available online at www.ehealthinitiative.org/news/survey_pressrelease.mspx.
2. Connecting for Health, Working Group on Accurately Linking Information for Health Care Quality and Safety. "Linking Health Care Information: Proposed Methods for Improving Care and Protecting Privacy." February 2005. Available online at www.connectingforhealth.org/assets/reports/linking_report_2_2005.pdf.
3. Ibid.
4. Gomatam, Shanti, Randy Carter, Mario Ariet, and Glenn Mitchell. "An Empirical Comparison of Record Linkage Procedures." *Statistics in Medicine* 21 (2002): 1485–96.
5. Connecting for Health. "Linking Health Care Information."
6. Ibid.
7. Montgomery Research. "RxHub: Where the Prescribing Industry Connects." Health Care Technology Project. July 17, 2004. Available online at www.hctproject.com/documents.asp?d_ID=2788.

Lorraine Fernandes (lfernandes@Initiatesystems.com) is senior vice president of healthcare practice, and **Michele O'Connor** is senior director of healthcare practice at Initiate Systems, Chicago, IL.