June 30, 2009

Privacy and Security Solutions for Interoperable Health Information Exchange

Perspectives on Patient Matching: Approaches, Findings, and Challenges

Prepared for

Jodi Daniel, JD, MPH, Director
Steven Posnack, MHS, MS, Policy Analyst
Office of Policy and Research
Office of the National Coordinator for Health IT
200 Independence Avenue, SW, Suite 729D
Washington, DC 20201

P. Jon White, MD, Director of Health IT
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Prepared by

Linda L. Dimitropoulos, PhD
RTI International
230 W Monroe, Suite 2100
Chicago, IL 60606

Contract Number 290-05-0015
RTI Project Number 0209825.000.015

Contract Number 290-05-0015
RTI Project Number 0209825.000.015

Privacy and Security Solutions for Interoperable Health Information Exchange

Perspectives on Patient Matching: Approaches, Findings, and Challenges

June 30, 2009

Prepared for

Jodi Daniel, JD, MPH, Director
Steven Posnack, MHS, MS, Policy Analyst
Office of Policy and Research
Office of the National Coordinator for Health IT
200 Independence Avenue, SW, Suite 729D
Washington, DC 20201

P. Jon White, MD, Director of Health IT
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Prepared by

Linda L. Dimitropoulos, PhD
RTI International
230 W Monroe, Suite 2100
Chicago, IL 60606

LIST OF AUTHORS

Shaun J. Grannis, MD, MS, Regenstrief Institute and Indiana University School of Medicine

Alison K. Banger, MPH, RTI International

David H. Harris, MPH, RTI International

ACKNOWLEDGMENTS

RTI gratefully acknowledges the members of the Technical Expert Panel for their thoughtful review and insightful comments:

Andrew Borthwick, PhD, Principal Scientist, Intelius, Inc.

Bradley Malin, PhD, Assistant Professor, Department of Biomedical Informatics, School of Medicine, Vanderbilt University

Scott Schumacher, PhD, Senior Vice President and Chief Scientist, Initiate Systems

Walter Sujansky, MD, PhD, President and Senior Consultant, Sujansky and Associates, LLC

Mari Tietze, PhD, RN-BC, Director, Nursing Research and Informatics, Dallas-Fort Worth Hospital Council Education and Research Foundation

CONTENTS

Section **Page**

Appendix

FIGURES

Number **Page**

TABLES

EXECUTIVE SUMMARY

As electronic health information exchange becomes more prevalent, the accurate and efficient matching of patients to their health records will become a greater and more pressing priority. Incorrect matching can result in misinformation and medical error and can compromise privacy and security if patient information is inappropriately disclosed. No standardized process to evaluate matching approaches currently exists, and only limited information is available about the performance of operational matching systems. The lack of a standardized method for matching is compounded by challenges such as patient information that is out of date or incorrectly recorded, the sharing of identifiers, and identity traits that are too common to allow an unequivocal match of records.

To address these challenges, the health care industry has proposed and pursued a number of approaches to patient matching, including deterministic (rules-based) matching, probabilistic (statistical) matching, biometrics (such as fingerprints or retinal scans), and the use of a unique patient identifier (UPI). Each of these approaches presents technical, logistical, and policy benefits, as well as concerns. Multiple factors may influence the development and implementation of matching solutions at the local or nationwide level, including adaptability, accuracy, scalability, sustainability, and privacy and security. Improved matching solutions should also include transparent evaluation, documentation, and dissemination. Successful matching of patients to their records requires research and input from both policy and technical experts.

Consequently, it may be time to revisit the use of a UPI for patient matching. Although this approach engenders significant debate, it has the potential to improve matching beyond the capabilities of algorithmic methods (e.g., reduce the number of false positive and false negative results) and consequently to improve clinical care. Additional policy mechanisms, such as strict legal requirements and heavy fines for those who misuse a UPI, may be required to make this type of solution a reality.

Matching patients to their records is a foundational component of electronic health information exchange. As the nation moves forward with the adoption of health information technology, the development of appropriate standards and policies for patient matching will be critical to ensuring quality clinical care and patient privacy.

1. INTRODUCTION

Established in June 2006 by RTI International through a contract with the U.S. Department of Health and Human Services (HHS), the Health Information Security and Privacy Collaboration (HISPC) has had overarching goals to assess variations in organization-level business practices, policies, and state laws that affect electronic health information exchange and to identify and propose practical ways to implement practices that will permit interoperability while preserving the privacy and security requirements set by local communities.

One such health care business practice is the methods by which individuals are uniquely matched to their health records—a foundational challenge to electronic health information exchange requiring both policy and technological solutions. The HISPC *Nationwide Summary* (Dimitropoulos, 2007c) reported that the standards and methods used to identify and match patients to their health information and health records varied widely. In addition, incorrectly matching a patient to a health record may have privacy, security, and health care implications, such as wrongful disclosure and treatment based on another patient's health information. The health care industry has proposed and pursued multiple solutions to address this challenge.

Health information organizations (HIOs)[1] are playing an ever increasing role in providing the ability to aggregate disparate sources of clinical data. However, recommendations that explicitly address HIO methods for patient matching are limited. Although organizations such as the Healthcare Information Technology Standards Panel (HITSP) disseminate health interoperability specifications that include patient identity management *transactions*, the specifications are silent with respect to patient matching *methodologies* and *algorithms* (Sloane & Carey, 2007; American National Standards Institute, 2008). Because HIOs represent complex "melting pots" of heterogeneous clinical information sources with varying data quality and characteristics, clear documentation and dissemination of concrete, real-world methods for accurate and efficient patient matching are crucial to improving electronic health information exchange.

Patient matching in an HIO setting is distinct from general record linkage with static databases because it requires real-time matching of very large databases without the luxury of additional review for every questionable match. In real-time matching, health care providers must be able to quickly receive accurate results with minimal manual review.

Incorrectly matching a patient's health record to the patient of interest (a false positive match) or failing to match a record to a patient that should be included (a false negative

[1] This report uses the term *HIO* as defined by the National Alliance for Health Information Technology (2008, p. 6): "HIO: An organization that oversees and governs the exchange of health-related information among organizations according to nationally recognized standards."

match) can have serious treatment consequences. Without a complete and accurate medical history, health care providers may be unaware of potentially fatal drug interactions when prescribing medicine or may mistakenly treat patients for a disease they do not have. Additionally, an incorrect match can lead to the disclosure of the wrong patient’s health information to health care providers or other persons.

This paper was developed to review the theoretical, experimental, and operational approaches of matching patients to their health records. In particular, we focus on HIOs and the privacy and security risks and benefits, ease of use, effectiveness, and scalability challenges of various patient matching approaches. The paper was reviewed by a Technical Expert Panel (TEP) to ensure quality and accuracy.

2. METHODOLOGY

We developed this paper using three main sources of information: a literature review, interviews with health information organizations (HIOs) currently conducting patient matching, and a review of reports previously published under the HISPC contract. These three components form the foundation of our analysis. We present the methodology for each component in this section.

2.1 Literature Review

Our literature review focused on articles that discuss approaches to patient matching in the health care sector. We located articles through PubMed searches, supplemented by sources from our previous research. Although a significant body of research exists on matching for other fields and industries, such as gaming, consumer research, and computer science, we considered these articles to be out of scope. The intricacies of the health care environment, including legal, technical, and policy facets, make it difficult to directly apply findings from other sectors to health care.

We include seminal articles germane to the fundamental principles of matching as references throughout the paper. Recent non–health care survey articles capturing the current state of the art in matching strategies and technologies that may be of interest to the initiated reader are included as an addendum to the overall literature review.

In conducting the literature review, we grouped articles generally into *technical* and *policy* subgroups. Technical articles were those that described a theoretical, experimental, or operational approach to patient matching. They usually, though not always, included a detailed description of the matching methodology or algorithm and analysis of the results (e.g., measures of sensitivity and specificity). The policy articles were those that took a broader view of the issue of patient matching by describing one or multiple options for matching patients to their records without delving into technical details or offering quantitative analysis. We divided the articles to determine whether technical approaches in use or being considered were concordant or discordant with proffered policy approaches. That is, were the policy-oriented articles advocating for approaches actually being used or proposed in the field?

Although these technical and policy article subgroups differed somewhat with respect to their detail, we evaluated them across similar dimensions:

Does the article describe theoretical, experimental, or operational approach(es) to patient matching?

- What are the privacy and security considerations of the approach(es)?
- What matching methodology, if any, was used?
- What is the effectiveness and/or accuracy of the approach(es)?

- What are the scalability and sustainability implications of the selected approach(es), including ease of use?

2.2 HIO Interviews

To supplement the information from the literature review, we contacted seven HIOs that conduct patient matching to understand their approach and methods. We divided this process into five steps:

- Identify operational HIOs.
- Develop a list of questions to understand how operational HIOs are approaching matching.
- Contact HIOs and determine if they are eligible to participate.
- Interview HIOs.
- Summarize and analyze results.

To determine which HIOs should be contacted, we compiled a list of operational HIOs from the lists of organizations that participated in the Nationwide Health Information Network (NHIN) Trial Implementations,[2] the Agency for Healthcare Research and Quality–funded State and Regional Demonstration projects,[3] and HIOs identified by the State-Level Health Information Exchange project as being in stage 3 (early implementation) or stage 4 (operating implementation).[4] We then developed a list of questions designed to give us a better understanding of how HIOs of various sizes are handling the challenge of matching patients with their records in a real-life setting (see Appendix A for the full list of questions).

After reviewing the list of questions for completeness and specificity, we contacted the HIOs to determine whether they were eligible to participate. Organizations were considered eligible if they were currently conducting patient matching or were planning to implement patient matching in the near future.

After initial contact, we sent the list of questions to the HIOs and asked them to respond by telephone or in writing. We scheduled and conducted interviews in April 2009. Some respondents provided answers during a telephone call; others wrote responses directly in the questionnaire. To ensure consistent results, we compared the responses obtained in the interviews with the written responses. Both approaches yielded the same quality and depth of information.

[2] The listing of the NHIN Trial Implementation participants can be found at http://healthit.hhs.gov/portal/server.pt?open=512&objID=1191&parentname=CommunityPage&parentid=9&mode=2&in_hi_userid=10732&cached=true

[3] The listing of the State and Regional Demonstration projects can be found at http://healthit.ahrq.gov/portal/server.pt?open=512&objID=654&&PageID=12043&mode=2&in_hi_userid=3882&cached=true

[4] The listing of State-Level Health Information Exchange initiatives can be found at http://www.slhie.org/efforts.asp

2.3 Information From Previous Stages of the HISPC Project

In the HISPC reports *Assessment of Variation and Analysis of Solutions* (Dimitropoulos, 2007a) and *Final Implementation Plans* (Dimitropoulos, 2007b), the state teams documented variations in privacy and security practices and developed solutions and implementation plans to address those variations.[5] Many states observed that matching patients to their records in a consistent, private, and secure way was essential to electronic health information exchange.

While these discussions were often at a higher level (e.g., discussing a general need for robust methodologies and technical standards), they indicate a real need for work in this area. This paper includes relevant information from the HISPC summary reports.

2.4 Technical Expert Panel Formation and Review

The Technical Expert Panel (TEP) was formed to provide expert review and input into the development of this paper. The members of the TEP have significant expertise both in the methodology behind matching patients to their records and in the implementation of matching systems. As experts in the field of patient matching, the TEP reviewed the paper to ensure that it accurately included applicable approaches to matching patients to their health information, ways the approaches vary, and the potential privacy and security risks and benefits to each approach.

The TEP reviewed a final draft of the paper and provided comments via a review form and notations in the text. We discussed their feedback at a conference call and, to the greatest extent possible, incorporated it into this final version.

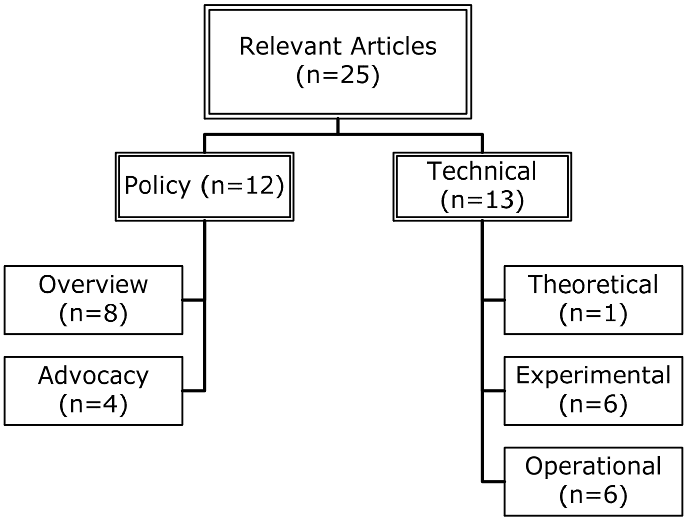
[5] The *Assessment of Variation and Analysis of Solutions, Final Implementation Plans,* and *Nationwide Summary* reports can be found at http://healthit.ahrq.gov/portal/server.pt?open=512&objID=654&&PageID=13062&mode=2&in_hi_userid=3882&cached=true

3. SUMMARY OF LITERATURE REVIEW

As noted in Section 2, we restricted our literature review to articles that discussed or applied approaches to patient matching in the health care industry. We include seminal articles from the field as references throughout the paper, but we do not specifically discuss them. In addition, several surveys of matching in other fields are referenced for readers interested in more details.

In reviewing the literature, we divided the articles into two sections: technical literature and policy literature. One specific goal was to determine whether solutions proposed by policy experts were consistent with technical solutions in use or under development. Figure 3-1 summarizes how the articles were sorted and evaluated.

Figure 3-1. Literature Review Results



3.1 Technical Literature Summary

In general, the peer-reviewed literature explored theoretical or experimental approaches to patient matching, rather than fully operational approaches. In addition, much of the literature focused on public health concerns (e.g., disease registries or prenatal health records), rather than matching patients to records for real-time clinical care.[6] There was a limited body of literature from U.S.-based research groups conducting matching in clinical settings. It may be that U.S.-based research groups are not likely to publish in peer-reviewed publications or that more of the work being done in the United States is

[6] The articles discussed here are Campbell, Deck, and Krupski (2008); Christen (2008); Dal Maso, Braga, and Franceschi (2001); Grannis, Overhage, and McDonald (2002, 2003); Karmel and Gibson (2007); Liu and Wen (1999); Lyons et al. (2009); Meray, Reitsma, Ravelli, and Bonsel (2007); Newman and Brown (1997); Pates et al. (2001); Sauleau, Paumier, and Buemi (2005); and Whalen, Pepitone, Graver, and Busch (2001).

proprietary and maintained as a trade secret. Alternatively, some organizations have expressed reluctance to disclose information about the performance of their matching system, for a variety of reasons.

The majority of approaches reviewed used probabilistic matching methods, although a few used deterministic or hybrid approaches. In the articles that compared probabilistic and deterministic matching approaches (e.g., Campbell et al., 2008; Whalen et al., 2001), probabilistic matching was found to be superior.

A disproportionate number of articles focused on matching in scenarios other than real-time clinical care. This finding suggests that the approaches described may need to be carefully evaluated for use in clinical care environments. A false positive or false negative match in a public health surveillance system poses a limited risk to how a specific individual is treated. However, in a clinical setting, an incorrect match could lead to inappropriate care or, worse, could harm the patient. Table 3-1 summarizes attributes of the technical literature.

Table 3-1. Technical Literature Matching Approaches

Attribute	Information From Technical Literature
Number of individuals/records reviewed	■ 26,000 to 500,000
Purpose of matching	—
Operational clinical	■ 2 of 13 (15%)
Theoretical clinical	■ 1 of 13 (7%)
Registries	■ 5 of 13 (38%)
Research	■ 5 of 13 (38%)
Country of origin	■ United States: 6 ■ Non–United States[a]: 7

[a]Australia (2), Canada, France, Italy, Netherlands, United Kingdom.

It is not clear whether approaches outlined in research literature can be readily adapted for use in real-time clinical care. The number of records or individuals examined is smaller than the number of matches that an HIO would be expected to perform, and the success of these algorithms is more accurately described as the success of the algorithm *plus manual review*. The smaller sample size (26,000 to 500,000 records compared with 225,000 to 9.4 million records for the HIOs) and the need for manual review could result in scalability issues if the approaches are implemented on a regional or nationwide scale where millions of individuals or records are involved. In addition, the algorithms are not compared against a standard data set, so it is not possible to compare the quality of results across the algorithms presented in the research. Generalizability is also an issue; a matching approach developed around a Dutch, Italian, or central Indiana population, for example, may not work well with different populations using alternate naming conventions. Given these limitations, matching approaches from research literature should be evaluated in a local context before being implemented.

3.2 Policy Literature Summary

Compared with the technical literature, which typically discussed the implementation of probabilistic versus deterministic matching and the algorithms involved, the policy literature included articles that offered a general overview of solutions and those that advocated for a specific solution.[7] What was most interesting was that the policy literature generally discussed algorithmic matching as a single category and mutually exclusive alternative to a unique patient identifier (UPI). In contrast, the technical articles considered the distinction between probabilistic and deterministic matching. Presenting algorithmic matching as the only alternative to a UPI establishes a false dichotomy and unnecessarily limits the scope of potential solutions.

There is a greater body of literature supporting the creation of a UPI, compared with algorithmic matching. Prior to the congressional action that prohibited the U.S. Department of Health and Human Services (HHS) from developing a UPI, HHS commissioned a study to review options for identifying patients. The study, conducted by Solomon Appavu in 1997, examined a wide range of options for matching patients to their health records against a detailed list of evaluation criteria. The options for identifiers were divided into three categories: unique patient identifiers, non-unique patient identifiers (identifiers specific only to the organizations that use them), and alternatives to unique identifiers. In all, 52 different traits across four categories were considered (30 conceptual characteristics developed by ASTM International, 5 operational characteristics, 6 component requirements, and 11 functional requirements) (Appavu, 1997).[8] A full listing can be found in Appavu's report, but examples of the requirements include such factors as assignability, accessibility, usability, whether the identifier could be used with existing infrastructure, and whether it can be used to accomplish care delivery and administrative functions.

In comparing each option against the evaluation criteria, Appavu found that only the UPI met all of the criteria. The other options (non-unique patient identifier and alternatives to a unique identifier) were deficient with respect to conceptual characteristics and consequently noncompliant with operational characteristics, component requirements, and functional requirements.

Hillestadt et al. (2008) noted that some form of unique identifier is needed to ensure accuracy in large databases. In examining an error-free database with 80 million records,

[7] The articles discussed here are Appavu (1997, 1999); Fernandes and O'Connor (2008); Greenberg and Ridgely (2008); Hillestadt et al. (2008); Markle Foundation (2005); Morrissey (2007); National eHealth Transition Authority (2006); Netter (2003); Rollins (2007); Stewart, Arellano, and Simborg (1984); and Wooster (2006).

[8] For an overview of Appavu's 1997 study, see Appavu (1999).

they found that it was necessary to use name, date of birth, zip code, and the last four digits of the social security number (SSN) to unambiguously match all records. Removing the last four digits of the SSN as a matching variable created nearly 1,000 false positives. Data in HIOs will not be error free, and so for a similar number of records, the number of false positives is likely to be higher than the estimates documented by Hillestadt et al.

A voluntary identifier administered by a nonprofit organization or a public-private partnership has also been proposed as a mechanism for implementing a UPI (Morrissey, 2007; Netter, 2003). Having a nonprofit entity or public-private partnership administer a voluntary identifier could ameliorate the public's concern of a government-run and government-maintained UPI. Use of voluntary identifiers is discussed in greater detail in Section 5.2.

In contrast, a 2005 report from the Markle Foundation recommended against pursuing a UPI and instead advocated for using probabilistic matching (Markle Foundation, 2005). The report documented four major reasons why a national identifier should not be pursued:

- It may be impossible to deploy a UPI system in the United States.
- Even if deploying a UPI system were possible, it would be delayed and potentially derailed by politically complex and sensitive issues.
- A national identifier presupposes successful solutions to other significant and presently unsolved technical challenges, including a method for "brokering trusted access to encryption keys" (Markle Foundation, 2005, p. 13) and developing a method for reworking existing systems to accommodate a new identifier.
- The expenses of such a system would be front-loaded, but the value postponed for years.

The Markle report also noted that the authors were initially skeptical that probabilistic matching methods would be sufficient on a nationwide scale, in which millions of transactions would occur. However, on the basis of current activity within the United States (both inside and outside of the health care environment), they concluded that probabilistic matching was feasible.

From the information currently available, it is difficult to fully evaluate the potential options for matching from purely a policy perspective. The cost/value element and privacy and security considerations for each matching approach have unknowns associated with them. From a cost/value perspective, cost estimates for implementing a UPI are available, but less information is available on the continuing costs to implement algorithmic matching. In addition, it is not clear where value accrues. For example, there may be diminishing marginal returns associated with moving from regional to nationwide exchange of data (Wooster, 2006). That is, the majority of value that results from linking records and sharing information may accrue at the regional level, and nationwide sharing may result in less gain at a higher cost. With respect to privacy and security, additional factors outside of matching, such as the system architecture, authentication, and access control policies, as well as enforcement, will affect how well privacy and security are maintained.

3.3 Other Literature Related to Matching

As noted in Section 2, we confined our literature review to articles that specifically discussed patient matching within the health care sector. However, there is a significant body of literature from other industries that may inform patient matching approaches for health care as more providers adopt electronic health record systems and as electronic health information exchange matures. For an overview of matching approaches used in other industries, see Winkler (2006) and Elmagarmid, Ipeirotis, and Verykios (2007).

4. CHALLENGES TO EFFECTIVE MATCHING

Health information is distributed across many independent electronic systems within and across organizations. A patient's health information may have multiple identifiers within a single institution or multiple identifiers across multiple institutions. This fragmented environment makes it difficult to correctly link information about individuals when it is needed for clinical and health care functions.

In the absence of a national unique patient identifier (UPI), common challenges to accurate patient matching include lack of consistent methods for conducting matching, out-of-date and incorrect information, recording errors, cross-cultural differences in naming conventions, identifiers that are too common to specifically identify individuals (e.g., the name "William Smith"), and identity theft and the sharing of identifiers. All approaches to matching operate within this complex milieu of factors and, as such, should be evaluated within this context.

4.1 Lack of Consistent Methods for Conducting Matching

HISPC states found that there is currently no consensus on patient matching accuracy thresholds or the method used to verify patient identifiers at the time of encounter. Each organization employs its own matching algorithm and patient matching methods, resulting in inconsistent results. States were concerned that without consensus on standards, an organization might send faulty data to a health information organization (HIO), which could lead to an incorrect match and potentially affect patient care. States also noted that a lack of trust between HIOs or states regarding the quality of patient information may hinder efforts to electronically exchange health information between the organizations.

4.2 Out-of-Date and Incorrect Information

Accurate matching of patients across different systems, such as hospitals or HIOs, depends on several critical factors. Organizations that connect to the HIO must provide sufficient, up-to-date information to allow for a match. Additionally, patients are under no obligation to inform providers when they move; therefore, a provider may not have the most current demographic information for a patient, making the matching process more difficult. Identifying traits that change over time are another challenge to accurate patient matching. For example, a person's last name can change with marriage, and geographic information changes when a person relocates. Further, family members sporadically share ostensibly unique identifiers, such as social security numbers (SSNs).

4.3 Recording Errors

Three kinds of recording errors are commonly introduced into health care data. First, phonetic recording errors occur with words and names that sound alike. For example, the first name "Shaun" can be recorded using at least three distinct patterns ("Sean," "Shawn," “Shaun”). Second, typographical recording errors occur when individual characters are inserted, deleted, or transposed. Third, morphologically similar characters can be mistaken for one another, particularly when transcribing handwriting or using optical character recognition; for example, the letters “I” and lowercase “L,” the numbers “1” and “7,” and the number “0” and the letter “O” can be interchanged.

4.4 Cross-Cultural Differences

Non-Western names can pose challenges when naming conventions differ. For example, Chinese names are typically written with the family name first. Thus, for Yao Ming, “Ming” should be entered as the first name and “Yao” as the last name. In addition, Hispanic and other cultures may have multiple family names. For example, the daughter of Ricardo González and Carmen Ramírez would have the name Rosa González Ramírez. Rosa’s last name should be entered as “González” or “González Ramírez,” but not as “Ramírez.” In either instance, the last name could be mistakenly identified during registration or when a staff member searches for a record.

4.5 Common Identifiers

Some cultures have a few last names that are very common; in Korea, five last names account for nearly 50% of the population. The lack of sufficiently discriminating identifiers and the inadequate number of identifiers also hinder patient matching because of false positives. For example, the Social Security Administration’s death master file contains more than 20,000 distinct males named “William Smith,” 5 of whom were born on January 20, 1920 (Social Security Administration, 2001). If name, gender, and date of birth are the only matching criteria used, these individuals may be falsely linked.

4.6 Identity Theft and Shared Identifiers

Identity theft and the sharing of documents, such as insurance cards, also pose challenges to matching. It is very difficult to detect medical identity theft or document sharing in a matching system unless a provider notices a discrepancy in the clinical data (e.g., some records indicate that a person has diabetes, while the others do not). The use of someone else’s information need not be criminal or malicious—it may be as simple as parents’ providing their identification credentials instead of their child’s. Systems must include a method for resolving such issues once they are identified, because they are difficult to prevent and detect.

5. POTENTIAL APPROACHES TO PATIENT MATCHING

Several approaches have been proposed to address the challenges involved in matching patients to their medical records. These approaches include the use of a unique patient identifier (UPI), a voluntary patient identifier, biometrics, and matching methodologies that are based on demographic information. This section provides an overview of these potential approaches to patient matching and the many components that can be implemented for each.

5.1 Unique Patient Identifier

One approach to patient matching proposes creation of a national UPI to identify, link, and locate records. If properly developed and implemented, a UPI could improve matching efficiency and accuracy. The cost of developing and deploying a national UPI has been estimated at between $4.9 billion and $12.2 billion, adjusted to 2009 dollars (Appavu, 1997; Hillestadt et al., 2008).[9] In addition, UPI implementation would likely take several years. Although a UPI could improve the matching process, it is not a panacea, for a variety of technical and political reasons.

From a technical perspective, supplemental patient matching methods would be needed when a UPI was absent, to identify duplicate patients within the UPI system and to accommodate historical data that have not been tagged with the UPI.

Political issues have limited the nationwide adoption of a UPI because of privacy and security concerns. As passed, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) calls for the creation of a UPI. The text of the law reads:

> (b) Unique health identifiers
>
> (1) In general
>
> The Secretary shall adopt standards providing for a standard unique health identifier for each *individual*, employer, health plan, and health care provider for use in the health care system. In carrying out the preceding sentence for each health plan and health care provider, the Secretary shall take into account multiple uses for identifiers and multiple locations and specialty classifications for health care providers.
>
> (2) Use of identifiers
>
> The standards adopted under paragraph (1) shall specify the purposes for which a unique health identifier may be used.[10] [emphasis added]

[9] The range presented in Appavu (1997) is $3.9 billion to $9.2 billion. Hillestadt et al. (2008) cite work done by the National Governors Association in 2006 calculating the cost of implementing a Real ID to be $11.1 billion. For comparison, these values were converted to 2009 dollars using the consumer price index.

[10] 42 U.S.C. 1320d-2(b)(1) et seq.

The idea of a UPI to be used across all institutions became a contentious issue, and in 1999 Public Law 105-277 prohibited the U.S. Department of Health and Human Services (HHS) from using any of its appropriated funds to develop a UPI without express congressional approval:

> None of the funds made available in this Act may be used to promulgate or adopt any final standard under section 1173(b) of the Social Security Act (42 U.S.C. 1320d–2(b)) providing for, or providing for the assignment of, a unique health identifier for an individual (except in an individual's capacity as an employer or a health care provider), until legislation is enacted specifically approving the standard.

The appropriations restriction has not been revisited by Congress, despite calls from some researchers and analysts (see, for example, Greenberg and Ridgely [2008]; Netter [2003]; Hillestadt et al. [2008]). Although states and other entities are not prohibited from developing and implementing a UPI, to date, no broad adoption or nonfederal promotion of a UPI has taken place in the United States.

5.2 Voluntary Patient Identifier

Another option for matching patients to their records is a voluntary patient identifier. A voluntary identifier could be administered by a nonprofit group or a public-private partnership and would mitigate some of the concerns related to a federally sponsored UPI, such as the existence of a government database that would hold extensive demographic information.[11]

One current example of a voluntary identifier to be used by health information organizations (HIOs) is being developed by Global Patient Identifiers, Inc. (GPII). GPII emerged as a result of the passage of two ASTM International standards (E1714 and E2553) that describe the architecture and implementation of the Voluntary Universal Healthcare Identifier (VUHID) system. ASTM established GPII to support and deploy the VUHID system.[12]

The VUHID system would not include a central database of patient information. Rather, it would work with HIOs' Enterprise Master Patient Index (EMPI) vendors, whose systems would in turn assign identifiers. The proposed process works as follows: A patient requests an identifier through a provider; that request is then routed to the EMPI and then to the VUHID system. The VUHID system validates the request and issues an identifier, which is routed back to the EMPI and associated with the patient's records. Finally, the identifier is returned to the provider and patient, and the patient receives an identification card.

GPII plans to offer two types of identifiers: an open voluntary identifier (OVID) associated with information that patients wish all of their health care providers to see, and a private

[11] See Netter (2003) and Wooster (2006) for additional theoretical discussion of a voluntary identifier.

[12] See http://www.gpii.info for additional information regarding GPII and the VUHID system.

voluntary identifier (PVID) associated with information that patients prefer to disclose only to selected providers. Patients can have multiple private identifiers, but only one open identifier. The private identifiers could be used to limit the disclosure of sensitive information, such as HIV or mental health information.

Although having multiple identifiers could help protect privacy, this approach does not take into account the ways in which administrative and clinical functions are currently linked. For example, if a patient sees a gynecologist and an internist at a hospital and wishes to keep the reproductive health records separate from the internal medicine records, it will be difficult for the hospital to produce a full billing record. In addition, allowing patients to have multiple identifiers assumes that patients are able to determine what information can be compartmentalized or withheld without endangering their health. Finally, using a voluntary identifier could create additional silos of information—information that includes the identifier and information that does not. A critical volume of information would likely need to be tagged with the identifier in order for the benefits of the identifier to be realized. It is not clear if creating these additional silos would perpetuate the challenges that already exist in matching or would remedy them.

GPII planned to announce its first beta test site in spring 2009, and it is not yet clear when outcomes from the beta testing will be available, although the results could offer evidence of the value of a voluntary identifier, the logistical challenges to implementation, and the privacy and security implications of using such an identifier with HIOs.

5.3 Biometrics

Biometric identifiers, including voice patterns, fingerprints, iris patterns, facial shapes, and vein patterns, are another method of uniquely identifying individuals in order to link data. The advantage of biometric identifiers is that they are highly specific to an individual, and identity can be verified without resorting to documents or cards that may be lost, stolen, forgotten, or altered. Disadvantages of biometric identifiers include the relatively expensive cost, fingerprint scanners notwithstanding. And although biometric identifiers generally remain stable over a person’s life, there are instances where the identifiers evolve. Voice patterns can change gradually with age or abruptly with illness, fingerprints can degrade (disappear) with time, and retinal patterns can change in patients with conditions that affect the eye, such as diabetes. Additionally, privacy concerns remain over the use of biometric identifiers for health care uses, because of the potential for biometrics, particularly fingerprints, to be used by law enforcement agencies (Prabhakar, Pankanti, & Jain, 2003).

5.4 Algorithmic Matching Approaches

Patient matching can also be accomplished using methods that depend on patient traits to match records. Patient traits may include unique identifiers, names, birth dates, addresses, sex, telephone numbers, and parents’ names. These algorithmic matching approaches are commonly described as being either deterministic or probabilistic. A third option involves combining both deterministic and probabilistic approaches (Campbell et al., 2008). The terms *deterministic* and *probabilistic* refer to the formal decision model that the approach implements. But these common labels fail to fully characterize matching algorithms because, beyond the decision model, algorithms include several key elements that affect matching performance. These elements include methods for measuring and ensuring data quality, candidate selection, and field comparison.

5.4.1 Measuring and Ensuring Data Quality

The quality of the data being matched strongly influences the accuracy that any given algorithm can achieve (Herzog, Scheuren, & Winkler, 2007). Data quality generally refers to data characteristics that influence the degree to which a specific data set can produce accurate matching results. Examples of characteristics that influence data quality (and ultimately matching results) include recording errors, missing values, and presence of highly discriminating fields. Although the performance characteristics of matching systems can be characterized in general terms, the specific performance of an algorithm should be evaluated in the context of real-world data.

Whether using probabilistic or deterministic matching methods, data-quality assessment typically occurs as an early step before a matching system is initially implemented and before incorporating a new data source into an operational HIO. Measuring data quality provides valuable information, and an initial analysis of the characteristics of patient matching informs the processes in at least two ways. First, it assesses how effectively each element in a data source can be used for matching and whether the data can fulfill specific, predefined matching-system performance requirements. Second, data-quality assessments can improve matching accuracy by identifying specific data shortcomings to be addressed by data cleanup strategies (data preprocessing). Examples of these strategies follow.

Standardizing Field Values

Fields within data sources must be formatted appropriately. Examples of standardization include parsing name fields into last name, first name, and middle initial. Date of birth can be parsed and standardized into individual components (year, month, and day) and also integer date from epoch.[13] Addresses can be parsed into components, most commonly street number and street name.

Identifying Invalid Field Values

Causes for invalid or erroneous values include default system placeholder values, typographical recording errors, phonetic recording errors, and deliberate misrepresentation. Methods for identifying invalid values include conducting frequency analysis of individual

[13] *Integer date from epoch* refers to a method of representing dates as a series of single integer values rather than in month/day/year format.

field values to look for outliers, comparing common values against collections of known common valid values, and applying simple validation rules to the data.

Identifying Invalid Values Using Term Frequency

Information systems often record a *default placeholder value* when either no value or an invalid value is provided. To identify such default placeholder values, which commonly lead to false positive matches, frequency counts for each field can be examined. Unusual values that are present with high frequency are identified as potentially invalid. Actual examples include name values of “SPECIMEN, LAB,” “DOE, JOHN,” “UKNOWN, NAME,” and date of birth “01/01/1900.”

Validating Field Values Using Rules

Values for certain person traits are governed by simple rules. For example, month of birth can assume 1 of 12 distinct values; day of birth can assume 1 of 31 values; and names can contain the letters A–Z, hyphens, and apostrophes, but no other punctuation and no numbers. Social security numbers (SSNs) and telephone numbers should not contain more than 6 nines or 6 zeros in a row. Person traits not complying with particular rules are marked as invalid and are not used for matching.

5.4.2 Candidate Pair Selection (Blocking)

The number of candidate pairs to be evaluated can become very large, because the number of pairs grows as a function of the Cartesian product of the records in the data sources to be matched. In other words, when linking one data set containing 1,000 records to a second data set containing 10,000 records, there are 1,000 x 10,000 = 10,000,000 total potential candidate pairs to be evaluated, most of which are non-matches. Because evaluating large numbers of pairs can hinder performance, particularly in a real-time matching system, approaches to overcome this problem have been developed.

Blocking is a strategy for reducing the number of candidate pairs by eliminating non-matches while retaining proportionately more true matches. This method is accomplished by restricting comparisons to those records that meet a given similarity measure; the approach is analogous to sorting socks by color before pairing them. Although many blocking strategies have been described, a common approach is to enforce simple exact-match agreement for different field combinations. (See Baxter, Christen, and Churches [2003]; Mohamed, Elfeky, Verykios, and Elmagarmid [2003]; or Gu, Baxter, Vickers, and Rainsford [2003] for additional discussion of blocking strategies.) For example, one blocking strategy may create candidate pairs for records that agree on full date of birth, while a second may create candidate pairs that agree on first and last names.

Characteristics of ideal blocking fields include high accuracy (few recording errors), high number of unique values, and uniform distribution (Jaro, 1997). Blocking approaches vary for different matching scenarios, depending on the quality of the data being matched and the performance requirements of the matching system. These approaches are typically designed to optimize the trade-offs between the computational cost of evaluating large numbers of records and the false negative rates caused by classifying pairs as a non-match, because of exclusion from the blocking partition. Multiple blocking strategies can be used to improve false negative rates while still reducing the overall number of candidate pairs (Jaro, 1997).

Blocking can improve system speed by reducing the overall number of candidate pairs to evaluate, but it comes with a cost of causing false negatives (missed matches). To mitigate false negatives, multiple blocking strategies can be used.

5.4.3 Field Comparison Methods

Matching algorithms compare corresponding fields between records to establish a pattern of agreement and disagreement among all fields. To minimize the effect of data variations (discussed in Section 4), field comparison methodologies, such as fuzzy match, allow inexact agreement between corresponding fields. For example, last names that agree on the first five characters but disagree on the subsequent characters may be declared to agree according to a fuzzy heuristic. Another example might be, "If birth date is within 1 month, then declare birth dates to be in agreement."

To loosen agreement requirements, phonetic transformation and string comparators functions may also be used. *Phonetic transformation functions* minimize the effects of recording errors among similar sounding words and include the Soundex, New York State Identification and Intelligence System (NYSIIS), metaphone, and double metaphone methods. The Soundex algorithm has five rules that transform words into an initial letter and a series of three digits representing consonants in the word. The NYSIIS algorithm has 11 basic rules that replace common pronunciation variations with standardized characters, remove repeated characters, and replace all vowels with the letter "A." Because the NYSIIS algorithm retains information on the sequence of vowels, it has higher discriminating power than Soundex. The NYSIIS transformations for "Shaun" and "Sean" are both "SAN." The metaphone and updated double metaphone transforms were developed to improve upon Soundex.

String comparators reflect agreement between corresponding fields by typically using a continuous metric ranging from 0 to 1, with 0 reflecting little or no agreement and 1 reflecting complete exact-match agreement. Examples of string comparators include the Levenshtein edit distance, the Jaro-Winkler comparator, and the longest common subsequence.[14]

[14] For additional information, see Christen (2006); Elmagarmid et al. (2007); Levenshtein (1966); Porter and Winkler (1999); and Sideli and Friedman (1991).

5.5 Decision Models

Once data are preprocessed, candidate pairs formed, and fields compared, then a decision must be made as to whether two records belong to a single patient. The role of the decision model is to assess the pair-wise agreement/disagreement patterns and adjudicate the pairs as a non-match, a match, or an indeterminate match requiring further review, often by a human.

5.5.1 Deterministic Models

Deterministic decision models may also be called *rule-based* or *heuristic* models. They typically identify matches by defining combinations of field agreement that are believed to reflect highly likely matches. Deterministic models often use exact match or other lightweight field comparators to establish field agreement. Once matching rules are established, the results are typically reviewed for accuracy before deployment. The accuracy of deterministic approaches is often highly dependent on the presence of discriminating identifiers (such as an SSN) or a local unique identifier (such as a medical record number). Deterministic rules tend to rely on the presence of highly specific identifiers and confirm matches with additional traits.

5.5.2 Probabilistic Models

Probabilistic decision models implement underlying statistical algorithms to identify matches. Examples include Bayesian algorithms, maximum entropy algorithms, and the Fellegi-Sunter maximum likelihood algorithm (Fellegi & Sunter, 1969). Probabilistic models produce accurate results by including configurable parameters that are customized to reflect characteristics of the actual data to be matched. Parameter customization is typically accomplished using one of several methods, including human review of sampled potential matches to estimate parameters, supervised bootstrap methods, and unsupervised parameter estimation methodologies such as expectation maximization (Winkler, 2000). Once parameters are configured, probabilistic decision models declare a link for candidate pairs with high match scores and declare a non-link if the score is very low. Candidate pairs that fall within an indeterminate middle range can be either declared a non-match or flagged for further human review.

Maximum entropy and Bayesian algorithms rely on training sets to estimate parameters for the underlying model. The disadvantages of these algorithms are that (a) estimating parameters can require costly human resources to manually review, and (b) these parameters may change over time. Therefore, these algorithms may require substantial resources to maintain accurate, up-to-date parameters specific to the data set of interest. Alternatively, the Fellegi-Sunter algorithm can use expectation maximization to accurately estimate parameters without a training set; it has been shown to be robust across broadly varying parameter combinations. Further, the Fellegi-Sunter algorithm provides accurate estimates for key matching metrics for any given matching scenario, including sensitivity and specificity for matches at any score threshold. Because of its ability to generate accurate, unsupervised estimates, the Fellegi-Sunter maximum likelihood algorithm is a core component of many probabilistic matching algorithms.

5.5.3 Combined Deterministic and Probabilistic Models

Recent analyses have studied the combination of probabilistic and deterministic methods (Campbell et al., 2008). A typical approach using this combination is to first identify matches among records with relatively high-quality data using deterministic methods. Remaining matches are then identified using a more sensitive (and forgiving) probabilistic algorithm.

5.5.4 Advantages and Disadvantages of Different Models

Each matching approach has strengths and weaknesses, and it is an open research question as to whether probabilistic or deterministic patient matching approaches are superior in the context of HIOs. Using simple rules, deterministic approaches can be implemented more rapidly than probabilistic methods. Deterministic algorithms tend to be straightforward and easy to comprehend, and computational requirements are typically minimal. Because deterministic methods rely on accurate and consistent data, they may not generalize well to other health care data sources with different data characteristics, depending on how the rules are selected and the nature of the variations in data. For example, a deterministic matching approach that includes the SSN as a matching field may not perform as expected in the context of a new HIO participating organization that does not maintain SSNs. To maintain matching performance and accommodate the change in underlying data characteristics, new deterministic decision rules may need to be constructed and validated.

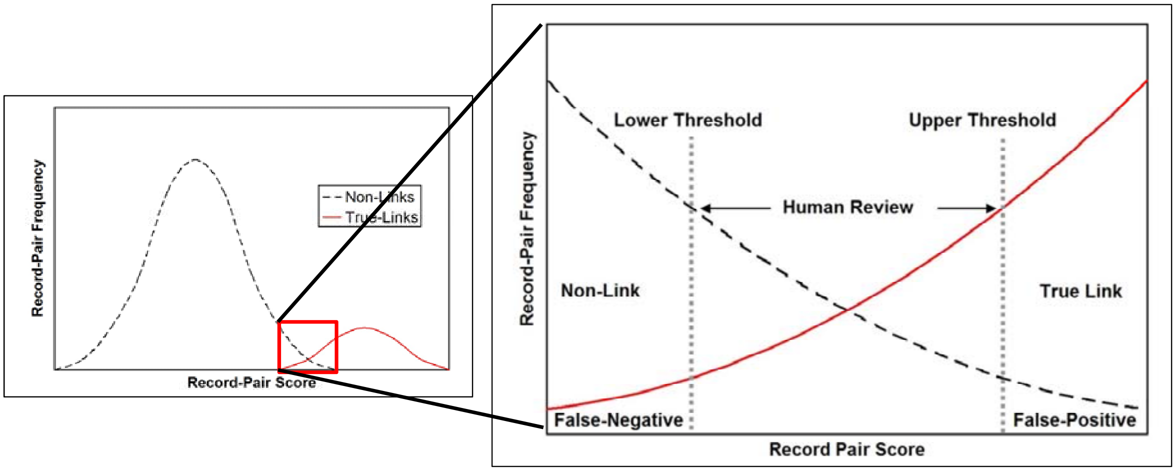
Alternatively, probabilistic approaches typically require specialized technical expertise to implement and maintain complex models. However, probabilistic models tend to be more forgiving of data errors and variation, and these algorithms can adapt to the data being linked by modifying the underlying model parameters. Applied to the situation just mentioned in which the SSN is missing, the probabilistic matching system parameters may need to be updated for the new data source; beyond that, it may require little or no change to the underlying decision model. The decision to modify or not modify the probabilistic approach assumes the availability of appropriate technical expertise.

5.5.5 An Illustration of Probabilistic Decision Models

To further illustrate the mechanics of probabilistic matching using the Fellegi-Sunter method, assume two data sets (file A and file B), each containing 10 records. Further assume that each record in file A truly links to a single record in file B, for a total of 10 true links. Further imagine that each record in file A is paired with all 10 records in file B by taking the Cartesian product of all possible combinations, resulting in 100 potential record pairs. Among the 10 true links, assume that one of the last names did not match exactly, either because of a recording error or because the last name changed or was unavailable. Therefore note that the last names exactly agree for 9 out of 10 pairs. This rate represents a 90% (9/10) agreement rate for last name among the true links. Further, assume that among the 90 non-linked pairs, the last name agreed by random chance for 2 of 90 pairs. This rate represents a 2% (2/90) agreement rate for last name among non-links. This component of the calculation, the chance of agreeing at random, is a function of the attribute value (e.g., ”Smith” vs. ”Einstein”), which leads to the observation that the weight for matching on “Smith” versus “Einstein” may vary. By taking the ratio of 90% divided by 2%, we get a quotient of 45, which suggests that record pairs agreeing on last name are 45 times more likely to be a true link than a true non-link. A similar process is carried out for all fields in the record pair. Weights for each field are combined to form a composite record-pair score. Field agreement contributes a positive weight, while field disagreement contributes a negative weight and reduces the overall record-pair score.

Using this model, the distribution of record-pair scores generally assumes a bimodal shape, with non-links receiving lower scores and true links receiving higher scores. There is often an overlap zone, or gray zone, where true matches and true non-matches exist in relatively equal proportions, as shown in Figure 5-1. Any record pair below a lower threshold is considered a non-link, and any record pair above an upper threshold is considered a true link, while those in the gray zone may be manually reviewed when human review is available. Minimizing the gray zone and creating efficient processes to disambiguate true and false links are important patient matching functions, because disambiguation requires costly human resources. When human review is not available, a single upper threshold is typically established. All matches below the threshold are treated as non-matches, while those above the threshold are considered true matches.

Figure 5-1. Illustration of the Intermediate Score Range Where Both True Matches and Non-Matches Are Present



NOTE: To disambiguate these linkages, human review is often necessary.

5.5.6 Clustering

Clustering refers to the process of evaluating relationships among records believed to belong to the same group. Relationships among a group of similar records may provide additional information that goes undetected in the typical pair-wise comparisons that form the basis of most matching systems. This additional information may help further identify true matches and non-matches. The unit of analysis for most matching systems is the single record pair. That is, most matching systems isolate each candidate pair to determine its match status. Once each pair is adjudicated, the relationships between similar records within a group can then be evaluated, and the aggregate group information may supply information that is undetected when records are compared in pair-wise fashion. For example, when a system establishes that record A appears to match record B, and record B appears to match record C, the pair-wise algorithm may produce incompatible results such that A does not match C (a potential false positive). This situation typically occurs among records with poor data quality (e.g., missing data, data recorded in error) and may not be uncovered until all three records are evaluated as a group. Various methods for adjudicating groups of pairs, including a process referred to as *transitive closure,* have been described (Cohen & Richman, 2002; Monge, 2000).

6. HIO MATCHING APPROACHES

To supplement the information gathered from our literature review and to gain insight into how patient matching is currently being conducted, we interviewed seven health information organizations (HIOs) from across the nation about their approaches to patient matching, including their consideration of privacy and security issues related to the matching process. (See Appendix A for a list of the specific survey questions.) The HIOs surveyed are in various stages of development. Although all HIOs interviewed have procedures in place for patient matching, one HIO has not yet begun to exchange data, one has been exchanging data on a pilot basis and will go into full production soon, and another recently went into production. This section describes some of the relevant characteristics of these HIOs. Table 6-1 presents a summary of their approaches to matching.

Table 6-1. Summary of HIO Matching Approaches

Attribute	Results
Number of unique patients in HIO	■ 225,000 to 9.4 million
Software type	■ Use commercial product: 5 HIOs ■ HIOs using commercial software that make some adjustments to the system: 3 HIOs ■ Use own matching solution: 2 HIOs
Type of matching	■ Probabilistic: 4 HIOs ■ Deterministic: 1 HIO ■ Combination of probabilistic and deterministic: 1 HIO ■ Fuzzy match based on heuristics: 1 HIO
Manual review component	■ 0 FTEs (full-time equivalents): 2 HIOs ■ 0.5 to 1.0 FTEs: 3 HIOs ■ Will use manual review, but number of FTEs TBD: 2 HIOs
Summary of all variables used by the HIOs for matching (note: not all HIOs use all of these methods)	■ Medical record number ■ First name ■ Middle name ■ Last name ■ Maiden or alias names ■ Gender ■ Date of birth (as single value or separate fields for month, day, and year) ■ Social security number (whole number or last four digits) ■ Phone numbers ■ Street address ■ City ■ County ■ State ■ Zip code ■ Driver's license number ■ Race ■ Marital status ■ Date of encounter

6.1 Size of HIO Populations, Users, and Activity

The number of unique patients contained in each interviewed HIO varies from approximately 200,000 to over 9 million. One HIO maintains a database that is primarily a push system to the doctor of record for clinical encounters, as opposed to a query-and-return system. This HIO reported that the push system processes approximately 80,000 electronic clinical encounter messages per day.

We also asked HIOs about their number of registration records. A patient may have multiple distinct registration records across participating organizations, sometimes referred to as *registration domains*. A registration domain refers to an entity that independently registers and tracks patient information for a variety of health care purposes. Most HIOs reported that they have between two and nine registration domains. However, several mentioned that a registration domain might be a health system that contains multiple hospitals and doctors' offices.

6.2 Matching Software and Types of Matching

Five HIOs reported using a commercial, off-the-shelf product for patient matching, while two have created their own patient matching solutions. The commercial products used by the HIOs are Initiate, Axolotl, Quadramed, and Sun eIndex.

In practice, the HIOs use a variety of methods to perform the linkages. One HIO uses only deterministic methods, another uses deterministic methods with the ability to use probabilistic methods for records with slight differences (based on the Markle Foundation [2005] guidelines for patient matching), while another uses fuzzy matching methods based on its experience performing matches. The other four HIOs use probabilistic methods with the ability to add deterministic rules. Several of the HIOs took the matching algorithm supplied by the software vendor and fine-tuned it for their particular patient population, with one HIO going so far as to supplement the program provided by the vendor with programs created in-house. One HIO mentioned that it purposely used the algorithm developed by its vendor because it was a small HIO and did not have the resources to create and maintain its own algorithm.

6.3 HIO Architecture

We asked HIOs to describe the general architecture of their patient matching method. All of the HIOs reported that they use a centralized master patient index (MPI) for the matching process and that the actual matching process is done centrally at the HIO. However, not all HIOs have a centralized database of patient data. Several reported that data are held in a federated manner—that is, each site provides the HIO access to its data but stores it locally on its own data servers. When a patient visits a provider, the provider communicates with the centralized MPI and sends information for the patient. The HIO then performs a match using the information in the MPI and decides if the patient is in the system. If the patient is in the system, the HIO pulls information from multiple sources and sends the provider one overall record for that patient. If the patient is not in the MPI database, the patient is added to the system with demographic information obtained from the visit.

6.4 Addressing False Positives and Manual Review

False positive matches are a concern for HIOs, because of the potential consequences of incorrectly matching a medical record to a patient. To help reduce the possibility of a false positive match, some HIOs manually review questionable matches. A questionable match occurs when there is not enough information to positively match patients to their records or when slight differences exist between the patient and the potential match. For instance, a patient may have recently moved and the HIO has not yet received the patient's new address. Or there may be a slight discrepancy in the date of birth, such as being off by one year (1941 vs. 1942, for example). By using other demographic information, the HIO can determine whether the two patients in question are indeed the same person.

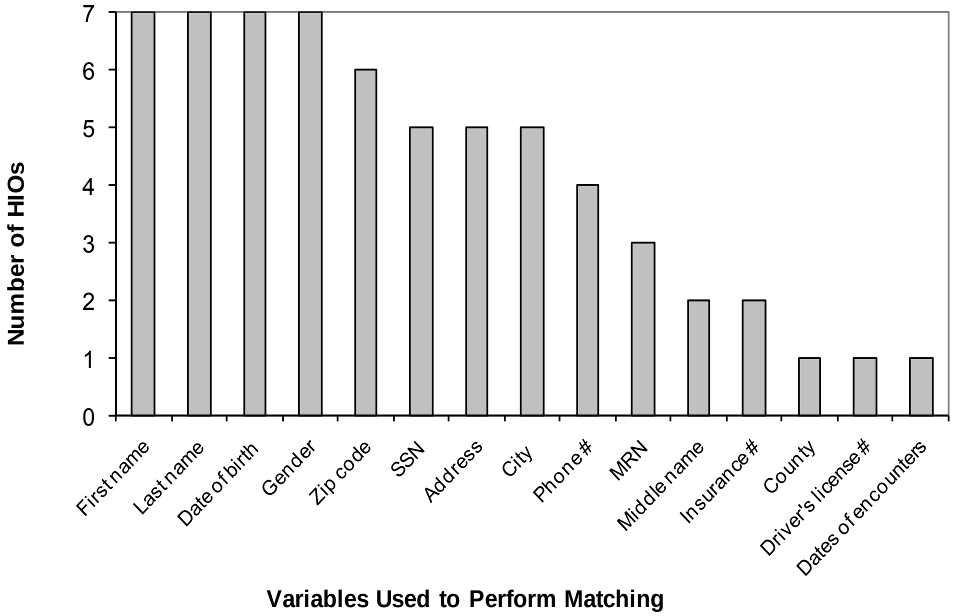
Some HIOs reported adjudicating questionable matches at the HIO level, while others put the responsibility for resolving the questionable match on the end users of the data, whether this is the registration clerk checking in a patient or a provider looking up records. Questionable matches are addressed through manual review of the records in question. HIOs reported using between 0.5 and 1.0 full-time equivalent (FTE) employees to perform the reviews. However, the 1.0 FTE hours may be split among several employees who perform the reviews as part of their overall job description. During the manual review process, the provider is presented with multiple possible matches if there is not an exact match and must determine which result, if any, is the correct one. In this case, to further protect patient privacy, demographic data for resolving a match is only exposed to end users for patients who have an existing care relationship with the organization performing the search. This practice can lead to the inadvertent disclosure of demographic information to end users for patients other than the person of interest. While this situation represents a risk, it is analogous to most clinical registration scenarios today: a patient's identity is clarified by providing information to the registration clerk, who reviews a list of potential matches typically based on last name or date of birth. The clerk makes a final selection only after verifying additional identifying fields—for example, "Mr. Smith, do you live on Elm Street?" Section 7 reviews methods used by HIOs to reduce inadvertent disclosure of personal information and to mitigate privacy concerns through the use of strong access control methods.

Regardless of the method used, as more records are added to the HIO database, the manual review of questionable matches will become a more time-consuming activity. Not all HIOs or providers will have the resources to devote more FTEs to the review process. Continuous fine-tuning of the matching algorithm according to the specific demographic make-up of the HIO database, as well as improving the quality of the data in the HIO database, can assist in reducing the number of questionable matches and therefore the number of false positives.

6.5 Variables Used to Complete Matching

To ensure that the correct records are matched with the correct person, HIOs use a wide variety of information to perform the match. Potential matching variables include first, middle, and last name; maiden name; gender; month, day, and year of birth; social security number (SSN); medical record number; address, including city, state, and zip code; phone number; race; driver's license number; and dates of encounter. HIOs reported that while the full SSN could be used in the matching process, either it would not be displayed in the results or only the last four digits would be displayed. The HIOs require between three and five of these variables for a match. All require first name, last name, and gender, with several requiring additional variables such as date of birth and zip code. One HIO reported that it will accept a match with only first name, last name, and gender; however, if one of these variables is not included, it requires a total of four variables. Figure 6-1 shows the frequency of the variables that the HIOs use to complete matches.

Figure 6-1. Frequency of Use of Matching Variables Across HIOs



NOTE: MRN = medical record number; SSN = social security number.

Although several HIOs reported using the medical record number or insurance (including Medicaid) number to assist in the matching process, none discussed the use of a unique patient identifier (UPI) as one of the variables used in matching. One HIO reported that what is generally thought of as a unique ID is not always static over time. This HIO uses an internal unique ID that is simply a group number attached to a collection of records that the linkage algorithm believes to be the same person. However, the HIO does not publish this unique ID, because it can change over time as more data about a person is gathered (e.g., if the HIO discovers that the person belongs to an existing group). In the absence of a national directive for a UPI, health care organizations generally use their own internal IDs, including medical record numbers, in the matching algorithm to group the patient's identity across organizations.

6.6 Setting False Positive Thresholds

As previously mentioned, false positives are a major concern during patient matches, because of their potential impact on treatment and privacy. To counter this, the HIOs choose to err on the side of false negatives during the matching process. A few of the HIOs reported that they have not yet established a false positive threshold but are in the process of reviewing identified false positive and other questionable matches in order to fine-tune their matching algorithm and minimize instances of false positives. Thus, they were unable to be more specific at the time of their interview. One HIO reported that its preferred threshold for false positives would be 0%, because it would rather miss a potential match than supply faulty data. However, while some mismatches can be readily identified by a reviewer, some are not detectable, making a goal of 0% false positives ideal but unattainable. Additionally, most of the HIOs either have, or are establishing, protocols to identify and adjudicate potential false positives.

6.7 Reviewing Twin and Familial Matches

Another issue facing HIOs is dealing with patients who have very similar demographic information, such as twins or other familial relationships (e.g., fathers and sons). This situation can lead to false positives, such as matching one sibling's data with the other sibling's name. Not only do twins have the same date of birth, they may also have SSNs that differ by only one digit, first names that are very similar (such as Ronald and Donald), and, in younger patients, the same home address. Solutions for this problem include fine-tuning the matching algorithm and manually reviewing these cases as potential matches. Requiring the first name to create a positive match will help safeguard against false positives among twins, while reviewing items such as the date of birth can assist in correctly matching nontwin family members.

6.8 Incorporating Feedback

To ensure that they have the most up-to-date and accurate information for a patient, most of the HIOs have a process in place to incorporate feedback from the registration domains and end users. One HIO reported that its federated approach allowed corrections and updates to be made to previously published or registered clinical documents, including corrections and updates to demographic and clinical information. The HIOs that have only recently begun exchanging data will continue to work on procedures for incorporating feedback.

6.9 Comparison of Real-World Approaches With Literature

The approaches described in the literature review and the real-world approaches used by HIOs to perform matching are only somewhat similar, as summarized in Table 6-2. The examples from the technical literature include smaller numbers of records, are generally aimed at public health or research uses, and use statistical or open source software. In contrast, HIOs are conducting real-time matching across a number of institutions and with larger numbers of records. Although two of the HIOs interviewed use their own matching software, the other five use commercial software that was not mentioned in any of the articles reviewed. This difference is likely the result of the different needs of researchers and clinicians, but it further elucidates the challenges in applying research strategies to real-time clinical matching.

Table 6-2. Comparison of HIO and Technical Literature Matching Approaches

Attribute	Information From HIO Interview[a]	Information From Technical Literature[b]
Number of individuals/records reviewed	■ 225,000 to 9.4 million	■ 26,000 to 500,000
Variables used for matching	■ Medical record number ■ First name ■ Last name ■ Gender ■ Date of birth (as single value or separate fields for month, day, and year) ■ Social security number (whole number or last four digits) ■ Phone numbers ■ Street address ■ City ■ County ■ State ■ Zip code ■ Driver’s license number ■ Race ■ Marital status ■ Date of encounter	■ First name ■ Last name ■ Gender ■ Date of birth (as single value or separate fields for month, day, and year) ■ Social security number ■ National identifier ■ Street address ■ City ■ Zip/Postal code ■ Date of encounter
Software used	■ Quadramed ■ Initiate ■ Axolotl ■ Sun eIndex	■ SAS (statistical software) ■ Automatch ■ Link Plus ■ Link King
Purpose of matching	—	—
Clinical	■ 7 of 7	■ 2 of 13 (15%)
Public health/research	■ NA	■ 11 of 13 (85%)

[a] Not all HIOs use all of these variables.

[b] Not all technical articles provided a complete list of all variables used to complete the matches.

7. ANALYSIS OF PATIENT MATCHING APPROACHES

Multiple factors affect the type of matching approaches that could be applied at the level of nationwide electronic health information exchange. This section outlines key issues associated with evaluating and implementing matching approaches: adaptability, false positive and false negative results, privacy and security, the ability to evaluate matching methods, accuracy, scalability, sustainability, ease of use, and inter-HIO exchange.

7.1 System Adaptability

Not all matching approaches may be successfully used in all circumstances. The context of matching imposes practical limits on the breadth of matching strategies that may be reasonably deployed. For example, in the context of a real-time, high-volume health information exchange, a matching system that requires human review of all indeterminate matches is infeasible, because of the high cost of human review. Health information organizations (HIOs) perform matching in different contexts that require different performance characteristics. Examples of varying use cases include high-volume, real-time matching tuned to high specificity; research-oriented batch matching tuned to high sensitivity; and de-identified matching tuned to optimize overall accuracy. This variability suggests that HIOs may need to support multiple matching strategies, because the same matching approach may not be optimal across all matching contexts.

The choice of matching approach is strongly influenced by human supervision and workflow timing constraints. *Human supervision* refers to whether manual resources are available for disambiguating uncertain matches, or whether matching will be largely performed in an automated fashion. *Workflow timing* refers to whether matching will be performed in a batch mode (where many records are evaluated as a set) or in real time (where each individual patient-level transaction is evaluated as it is received). Strategies for matching use cases may fall into one of four general categories on the basis of these constraints, as illustrated in Figure 7-1. Although these constraints are reflected in a dichotomous fashion, in many cases a continuum exists along these dimensions.

Batch mode matching with human supervision is often used in research and to create aggregate reports when high accuracy is required. HIOs may not want to use a human operator, either because of the high cost (human review can require thousands of person-hours of work) or because of privacy concerns. In such a case, HIOs may choose to automatically link patient data in real time without human supervision. De-identified matching is one example of unsupervised batch mode matching. To preserve privacy, identifiers may be encrypted or replaced with randomly generated tokens, which eliminates the ability to perform human review. An example of real-time matching with human supervision is its use in most large hospital systems, where many thousands of patient transactions are aggregated daily, and human operators disambiguate indeterminate matches. Human reviewers are allocated for disambiguation in this case because hospitals derive value from accurate patient data, which are used for revenue-generating business processes such as billing and invoicing. These distinctions are important, because much record-linkage research has studied modest size data sets as a one-time batch effort, which is different from HIOs, where matching is done in the context of an ongoing, dynamic, and large system.

Figure 7-1. Examples of Matching Scenarios Broken Down By Dimensions of Workflow Timing and Human Supervision

		Workflow Timing	
		Batch Mode	Real Time
Human Supervision	Substantial Manual Supervision	**Reporting, Research**	**Health Care Enterprise (Hospitals)**
	Little or No Manual Supervision	**De-identified Matching**	**Health Information Exchange**

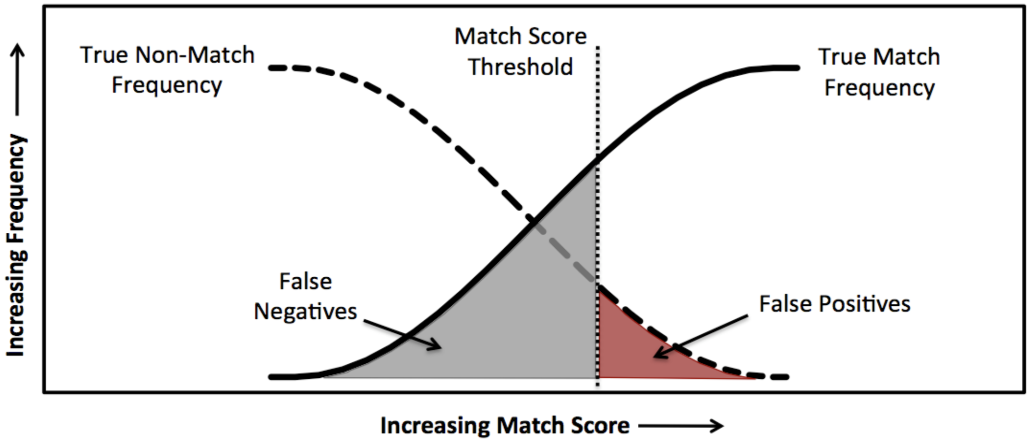
7.2 False Positives and False Negatives

When considering the performance characteristics of a matching system, it is important to understand the tradeoff between false positive and false negative matches. A false positive match occurs when two truly non-matching records are declared to match, while a false negative match occurs when two truly matching records are declared to be a non-match. Matching systems are often tuned to minimize either false positive or false negative matches, depending on which error is deemed to produce a more undesirable event.

For example, health care matching systems are generally tuned to minimize false positive matches, because a false positive match is generally considered to be more undesirable than a false negative. When records for two separate patients are merged (a false match), it poses a risk for inadvertent disclosure of protected health information and may lead to incorrect treatment decisions (e.g., treatment based on a nonexistent condition). On the other hand, false negatives limit privacy risks, because no information is disclosed, but they too may lead to incorrect treatment decisions (e.g., giving a patient a drug to which he or she is allergic, because the information was unavailable). However, it is worth noting that false negatives are, effectively, the status quo in health care—information is spread across a variety of paper and electronic systems, and physicians frequently lack a full view of a patient's medical history and health records.

Although eliminating all false positives may be desirable, it is impractical because false positive and false negatives are inversely related. That is, as the false positive rate is reduced, the false negative rate increases (see Figure 7-2). Low false positive rates are paid for by higher false negative rates: to achieve the ideal false positive rate of zero, the false negative rate approaches 100%, resulting in an ineffective matching system. Consequently, it is important that matching-system implementers understand the relationship between false positive and false negative matches for their specific target data. This understanding is necessary to make an informed decision regarding (a) whether their data can support the desired performance characteristics, and (b) at which specific choice of false positive and false negative rates to operate their matching system.

Figure 7-2. Illustration of the Relationship Between False Positive and False Negative Matches



NOTE: As the match score threshold is increased, the number of false positives decreases, but false negatives increase. As the match score threshold is lowered, the number of false negatives decreases, but false positives increase.

7.3 Privacy and Security Considerations

In the context of patient matching, three key privacy and security issues to consider are (1) the potential for inappropriate disclosure of health information due to false positive matches; (2) the simple receipt of an affirmative response that a record or patient information is available at a location that the patient may otherwise want to remain private (e.g., a record is matched at a center that primarily treats HIV patients); and (3) the repeated transmission of demographic data in instances where algorithmic matching is being used, since such information, when attained, generally includes enough personal attributes to commit identity theft.

All of the HIOs interviewed expressed concern over the privacy and security of patient data and have taken steps to safeguard this information during the matching process. HIOs reported using a variety of methods to ensure patient privacy, such as using role-based access protocols to ensure that only users that are authorized and have a need to view patient records are granted access to the matching software.

When matching patients to health records, there is always the danger of matching one patient (the patient of interest) with another's record. This error is known as a false positive match and, although uncommon, can have serious consequences. There are three potential outcomes from a false positive:

1. Treatment is altered, but no harm is done. For example, a physician mistakenly concludes that a patient is allergic to penicillin and gives the patient a different antibiotic, such as erythromycin, to treat sinusitis.
2. Treatment is provided, and harm is done. For example, diabetic patient A takes pills only for diabetes, while diabetic patient B takes 50 units of long-acting injectable insulin daily. Patient B's data (including the 50 units of long-acting insulin daily) are falsely linked to patient A. Patient A is admitted to the hospital for surgery, and the physician prescribes long-acting injectable insulin for patient A post-operatively. Patient A suffers a diabetic coma due to severe hypoglycemia.
3. Treatment is withheld, and harm is done. Great harm is a possibility in cases where few (perhaps one) treatment options are available, and the false information suggests that the patient cannot receive the treatment. For example, assume two elderly twin brothers, Ronald and Donald, who live in an area served by an HIO. Ronald suffers a hemorrhagic stroke. Two months after Ronald's stroke, Donald arrives at the emergency room with an ischemic stroke that has left him unable to talk. If the matching system falsely links the twins, and the emergency room physician thinks it was Donald who had the hemorrhagic stroke, the physician will withhold thrombolytic therapy to treat the ischemic stroke, because it would be contraindicated for a person who has recently had a hemorrhagic stroke. As a result of not receiving thrombolytic therapy, Donald may develop permanent neurological deficits.

The chances of a false positive causing serious harm are rare—it requires that the false positive occur and that the incorrect information makes sense. For example, another contraindication for thrombolytic therapy is recent heart surgery. However, if the records indicate that a person had heart surgery 5 days prior and a physical exam did not reveal any healing wounds or other physical signs of surgery, a physician or nurse would likely question whether the person truly had surgery.

Through the process of reducing false positive matches, the HIOs create a greater number of false negative matches, therefore missing clinical encounters that really should have matched to the patient of interest and become part of the patient's medical record. This situation creates a trade-off that, while increasing privacy and security by preventing the inadvertent disclosure of data, also potentially prevents a provider from having a full and complete medical record for a patient.

The privacy and security concerns related to false positive matches deal not only with inadvertent disclosure of data by matching one patient's data with another patient's medical record, but also with the process used to adjudicate questionable matches. To help reduce the number of incorrect matches and increase the number of correct matches, either HIOs or providers manually review questionable matches—those that fall between the true match and true non-match thresholds set by the algorithm.

One HIO reported that potential matches are constrained if the search parameters produce 10 or more possible matches for the search. The matching program instructs users to enter additional demographic information to bring the potential matches to fewer than 10 before any records are displayed. If no additional demographic information is available, users are not able to complete the matching process. In this system, the potential for the inadvertent disclosure of patient data could be a privacy concern. If the match were to occur on a national level instead of an HIO level, the number of instances where a match cannot be found because of lack of sufficient information may increase. This issue illustrates the need for accurate, up-to-date, and complete information, as well as the need for scalable and sensitive matching solutions.

Most HIOs reported that during the matching process, the system can potentially return more patient information than was originally entered for the match. The information returned varies by HIO, but it is always demographic data (such as name, address, and date of birth), not information relating to medical treatment or diagnoses. Because of privacy and security concerns, including the potential for identity theft, the social security number (SSN) is generally not displayed; in the cases where it is displayed, only the last four digits are shown. One HIO reported that although the address is not shown on the screen, when the user moves the mouse over the patient record, the address is shown on the screen. This protocol helps ensure patient privacy while assisting with questionable matches.

Another potential privacy concern related to patient matching occurs when a record or information about an individual is indicated as an affirmative match but is restricted in some way or requires the person's express permission to be disclosed. In this situation, even the acknowledgment that the patient has information at a particular location (e.g., a substance abuse treatment center) could convey private information. One HIO reported that patients can select which providers can access their records. Therefore, providers can view results only when the patients have permitted them to view their health information. Several HIOs also reported that in cases where the provider does not have permission to view the record, the matching program returns a response of "patient not found," rather than indicating that records exist but that the provider does not have the appropriate permissions to access them.

A final privacy and security issue related to patient matching involves the transmission of large batches of personally identifiable information, such as name, date of birth, address, and in some cases SSN, to complete the matching process. This information could be intercepted and stolen during transmission and used for identity theft. An argument has been made that the use of a unique patient identifier (UPI) could help to improve the privacy of patient data by minimizing the need to transmit other, more sensitive, information. Although the UPI could be stolen during transmission, it could be used only to steal medical record information, as opposed to financial data (Greenberg & Ridgely, 2008).

7.4 Evaluating Matching Methods

Patient matching methods are typically evaluated using one of two approaches. First, the output of an algorithm using real-world data can be compared against manually reviewed, discrete entity-level linkages, with each potential match determined to be a true match, a true non-match, or an uncertain match. Because the large number of records usually makes full manual review of all linkages infeasible, manual review is typically performed on a fractional sampling of the data, and statistical inferences are made. Alternatively, synthetic data can be used to create an a priori “gold standard” against which the algorithm’s performance can be easily measured. Although this approach is attractive because it may obviate the need for manual review, the validity of synthetic evaluations is strongly dependent on the degree to which the synthetic data reflect the underlying characteristics of the targeted real-world data (Christen & Pudjijono, 2009). Synthetic data are thus useful for debugging software, but less useful for performance evaluation. The limitations of these evaluation methods must be considered when they are applied.

7.5 Matching Accuracy

Because HIE stakeholders hold expectations of high accuracy for matching systems, metrics to evaluate accuracy are needed. *Accuracy* often refers to a general set of metrics characterizing the degree to which a system correctly identifies true matches and true non-matches. Accuracy has a specific mathematical definition:

$$\text{Accuracy} = \frac{\text{TM} + \text{TNM}}{\text{TM} + \text{TNM} + \text{FM} + \text{FNM}}$$

where TM = true match count; TNM = true non-match count; FM = false match count; and FNM = false non-match count.

In addition to accuracy, other useful metrics for assessing the degree to which a system correctly identifies true matches and true non-matches include sensitivity, also known as recall (sensitivity = TM/(TM + FNM); positive predictive value, also known as precision (PPV = TM/(TM + FM)), and specificity (specificity = TNM/(TNM + FM)).

Accuracy metrics are typically estimated empirically for a given real-world data set. As mentioned previously, matching systems can generate large numbers of pairs, so the output of a matching system is fractionally sampled, and human review establishes the rates for true matches, true non-matches, false matches, and false non-matches. Although human review is often considered the gold standard for evaluating matching performance, missing or inaccurate data can lead reviewers to disagree over the match status of records, which can lead to inconsistent manual review results. As a result, manual review analyses may benefit from including an indeterminate match status classification for records with poor data quality.

Because the performance characteristics of a matching algorithm are strongly influenced by the quality of the actual data, HIOs should consider the following framework for evaluating performance:

1. Develop a process to establish requirements for the level of accuracy needed within their system.
2. Evaluate the quality of their data and assess which matching approach (if any) can support their requirements.
3. Implement procedures for ongoing monitoring of data quality and matching performance.
4. Establish processes to address the specific issues if quality or matching performance fail to meet minimum requirements.
5. Establish policies for adjudicating disputed matching results.

7.6 Scalability

Scalability is a property that characterizes a system’s potential to either accommodate increasing amounts of work in a graceful manner or be readily enlarged (Bondi, 2000). In the context of record matching, at least four key factors can affect scalability of a matching system:

- the total number of distinct identity instances within the scope of the system
- the volume of transactions to be adjudicated per unit time
- the computational requirements of a given matching algorithm
- the architecture of the system

The first factor is the total number of distinct identity instances within the system. In the current context, *identity instances* equates to a patient registration within a given registration domain; the search space increases as the number of distinct patient registrations captured in an HIO increases. As the search space increases, increasing numbers of potential matches must be evaluated, requiring greater computational resources. Consequently, matching systems for HIOs should be designed at the outset with the notion that the patient population may increase dramatically from initial estimates, and the system should implement highly efficient strategies for identifying candidate matches.

A second factor that can affect scalability is the volume of transactions to be adjudicated per unit time. HIOs may need to adjudicate (match) hundreds of thousands to millions of new clinical transactions per day, and this rate may vary in both predictable and unpredictable ways. For example, it is well-known and predictable that far more clinical transactions are generated on weekdays than on weekends. Further, as an increasing number of participants connect to an HIO, the likelihood that one of the HIO participants will change a registration system increases. When a registration system is changed, it is not uncommon for systems to sporadically generate large volumes of merge transactions in a short period of time. One HIO surveyed experienced a daily surge of nearly 1 million registration transactions over a period of 2 days from a *single* registration domain. Additionally, when new participants join an HIO, the existing patient identities from that participant must be loaded and matched into the HIO's entity identification system; this process can place substantial additional burden on the system. Consequently, HIO entity identification systems must scale to accommodate expected and unexpected variations in clinical transaction volumes.

A third factor that can affect scalability is the matching algorithm itself. As a general rule, deterministic or rule-based algorithms tend to implement less sophisticated and less computationally intensive matching strategies. Matching algorithms that are based on an underlying statistical model can require more computing power and more sophistication on the part of the system implementer. The performance characteristics of the matching algorithm should be well understood and well characterized in the context of a given HIO prior to implementation.

A fourth factor that can affect system scalability is system architecture. The two primary architectures discussed in the context of patient matching in HIOs are peer-to-peer and centralized matching. Peer-to-peer matching approaches, which require collecting patient identity information from multiple source systems with varying designs, face potential scalability bottlenecks in two areas. First, any network delay when communicating identity information between two systems will slow the matching process. Second, translating identity information from different systems in a variety of formats is inefficient and poses performance challenges. Most HIOs surveyed plan to implement, or have already implemented, a centralized index of patient demographic data. Although no decentralized HIO matching architecture has shown success and the trend appears to be toward centralized matching, this current situation does not imply that a decentralized patient matching architecture will not work eventually.

7.7 Sustainability

Many HIOs are in the process of identifying strategies for sustaining their operations. Many HIO implementers are finding that several layered *value propositions,* or services, are needed to sustain HIOs (Kansky, 2007). The ability to link patient-level data across multiple systems is a crucial component for any HIO service requiring patient-centric data to be joined across all HIO participants. Those services include delivering a just-in-time patient-centric summary of recent clinical data at the point of care and providing a comprehensive, longitudinal, aggregate patient-centric query tool for clinicians. And while patient-centric aggregation of clinical data is not an absolute functional requirement of HIOs, correctly matching patients to their records is necessary to fully realize the value of mobilizing health care data (Walker et al., 2005).

One key factor related to patient matching that may influence the sustainability of the health information exchange is the ongoing costs required to maintain matching functionality in an HIO. Those costs include human capital, software licensing fees, and hardware costs. Human capital is necessary when the system requires manual disambiguation of uncertain matches. In our survey, some HIOs use routine human review to disambiguate uncertain matches, but not all patient matching approaches implemented in HIOs require routine human review. For those organizations that use a vendor solution for patient matching, ongoing licensing fees will factor into their costs. A typically lesser but ongoing expense is the cost of the hardware necessary to operate the patient matching functionality.

7.8 Ease of Use

Implementing a patient matching solution is a complex analytical process that typically requires subject matter expertise in matching. Much of the complexity of patient matching lies in understanding the nature of the specific data to be matched. Therefore, analytical tools that help assess the unique characteristics and quality of the specific data to be matched can aid implementation of a matching strategy.

These tools improve the efficiency and ease of creating a matching strategy by highlighting field-specific data characteristics that can (positively or negatively) affect the matching process. Such characteristics include rates of missing values, distribution of common values, presence of invalid values, dependencies between fields, and average frequency of unique values. This information guides selection of specific data preprocessing steps and helps identify optimal fields to be used for selecting candidate matches.

After a matching strategy is implemented with the assistance of analytic tools, and when manual disambiguation of uncertain matches is required, ongoing support and maintenance of the patient matching system can be augmented with tools that help end users identify potential false positive and false negative matches. As is the case most commonly with probabilistic matching approaches, candidate matches are assigned a score. Low scores are typically labeled non-matches, high scores are typically labeled true matches, and intermediate scores are indeterminate and can be flagged for manual review. Such manual review tools are typically customized for the particular matching approach.

7.9 Inter-HIO Patient Matching

This paper addresses patient matching issues largely from the perspective of a single HIO (*intra*-HIO matching). As the number of HIOs grows, the need for *inter*-HIO data exchange will arise. With that need will come the necessity to examine new technical and policy-related issues pertaining to inter-HIO data sharing and patient matching. The Connecting for Health Common Framework describes initial perspectives on a record locator service and explores technical and policy issues related to inter-HIO information exchange and patient matching (Markle Foundation, 2006a, 2006b).

One potential issue to explore is the impact of exchanging data between HIOs that operate matching systems with different operating characteristics. Each HIO will establish requirements for its internal matching performance characteristics and must ensure that its data and matching system can support those requirements. The particular operating characteristics will likely vary among HIOs, either by design or because of data quality: one may have a false positive match rate of 1 per 100,000 matches, while another achieves a lower 1 per 1,000,000 rate. Although it may be infeasible to change operating characteristics for HIO patient matching systems, it may be useful for HIOs to understand the nature of the data being exchanged and the relative risk of a false positive.

Another related issue that may lead to false positives is the potential variance in data fields used for matching, varying population characteristics, and lack of shared knowledge related to these variances between HIOs. For example, “Eduard Chavez” may be a common name in HIO A and an uncommon name in HIO B. HIO A permits use of the SSN for matching, but HIO B does not. HIO A queries HIO B for “Eduard Chavez.” HIO B’s matching system, believing “Eduard Chavez” is a unique entity, returns all data for Mr. Chavez, who happens to be the wrong individual, resulting in a false positive. Allowing HIOs to exchange aggregate summary statistics that characterize their respective populations may mitigate this scenario. Further, since matching algorithms are often tuned to the specific data characteristic of a given system, if two or more HIOs begin to match patients between systems, those systems may improve matching accuracy by sharing aggregate statistics, which parameterize their algorithms.

These and other broader issues, such as differing technologies, policies, and architectures, suggest that further analysis of inter-HIO data exchange may be necessary.

8. CONCLUSIONS

This paper represents an introductory overview of the current state of patient matching in the context of health information organizations (HIOs), functional and technical considerations, and associated privacy and security issues. Most HIOs interviewed are in the nascent stages of implementing patient matching and are pursuing a variety of approaches. Formal documentation and review of successful best practices and lessons learned related to this complex process will likely emerge as the implementation process progresses. System monitoring is vital: although most HIOs described the ability to log matching decisions, processes for ongoing evaluation and monitoring of matching performance are in the early stages.

Our conclusions fall into two general categories. The first relates to the need for transparent evaluation, documentation, and dissemination of the functional aspects of matching in the context of HIOs. The second concerns conceptual and policy-related matters, such as privacy and security of patient information, in the context of matching.

8.1 Functional Aspects and Transparency

Conclusion 1. A framework for describing detailed approaches to matching, including technology, human resources, and workflow, is needed. Although many publications have demonstrated accurate patient matching under specific constrained circumstances, further work is needed for several reasons. A chief function of emerging HIOs is to aggregate health information from many sources that themselves gather health information using different business and data validation processes, apply different default codes for missing values, and collect different data traits, among other differences. These variations pose unique challenges for matching patients to their health records. In the face of these challenges, there are remarkably few descriptions of system approaches to matching in an electronic health information exchange environment. Creating a descriptive framework for disseminating different matching approaches will help convey current best practices. Thus, it may be beneficial to create a consistent framework for HIOs to characterize different matching system technologies, degree to which human review is involved in adjudicating matches, and impact on clinical workflows.

Conclusion 2. Consistent approaches to evaluating and disseminating the accuracy of various matching strategies, including those that employ human review, are needed. The attainable level of accuracy in the setting of an HIO is currently largely undefined. It is also unclear whether a single algorithmic approach to matching (e.g., deterministic or probabilistic) is superior in the context of HIOs or whether multiple approaches can equally satisfy HIO requirements for system performance, accuracy, scalability, ease of use, and flexibility. Further, many matching algorithms that have proven successful are the result of the algorithm plus human review for disambiguation (Pates et al., 2001), and matching systems often incorporate human review. Also, since much of the projected savings related to widespread electronic health information exchange are derived from improved efficiency, requiring human review for disambiguation of uncertain record matches will adversely affect cost savings (Shekelle, Morton, & Keeler, 2006). Few studies have explicitly evaluated automated matching with no or little human review in the context of HIOs, nor has the cost of human review in the context of HIOs been evaluated. Therefore, specific formal evaluation of the performance characteristics of different matching approaches can help inform the state of the art. Further formal exploration of the degree to which HIOs are implementing human review for matching and the degree to which this strategy influences matching system performance is warranted.

Any framework should attempt to evaluate and describe the false positive and false negative rates achieved by the matching system. It is important that system implementers, HIO stakeholders, and policy makers understand the vital principle that perfection is unattainable for most real-world matching systems. False positive and false negative rates are inversely related, and at least one (and typically both) will be non-zero. The privacy and security implications of carrying degrees of imperfection should be addressed for each HIO. The combination of these rates is sometimes referred to as the *operating envelope*.

Conclusion 3. Further exploration of inter-HIO matching will be warranted as the need for data exchange between HIOs increases. First, it is unclear whether the operating envelope for matching systems within different HIOs will be similar or will differ substantially. Second, matching between HIOs increases the total population being searched, and false positives become more likely in larger patient populations. If operating envelopes differ or if population size substantially alters performance characteristics of the HIO matching systems, then exploring the implications of these differences will be informative. If algorithmic matching approaches cannot meet requirements across larger systems, then the need for a universal patient identifier (UPI) may become clearer. There are currently no formal evaluations of the matching interface between HIOs and the impact that matching system differences (*impedance mismatches*) may have on matching performance.

8.2 Conceptual and Policy-Related Matters

Conclusion 1. As health information technology progresses and health care stakeholders become more accustomed to the notion of mobilizing health care data, the concept of a UPI (voluntary or otherwise) may need to be revisited, provided that appropriate safeguards are in place, such as strict legal prohibitions related to its use beyond matching. However, it is important to recognize the caveats associated with such an approach. It is not a panacea: algorithmic approaches will continue to be necessary, because patients will not always have the identifier available and maintaining an unduplicated master list of UPIs matched to patients cannot use the UPI itself; roll out will take substantial time (potentially years), so the benefit will accrue gradually; and the transition will be complex and costly.

Conclusion 2. Further exploration and framing of HIO processes that expose demographic data for disambiguating indeterminate matches can inform ongoing HIO development and implementation. Although it may not be ideal to allow users to determine which match is the correct one, from a practical and workflow perspective it may be unavoidable. The patient lookup process by its nature exposes demographic data to end users and, in the case of false positive matches, may inadvertently disclose clinical information as well. To limit demographic exposure in existing matching and patient lookup approaches, policies and access controls should be used to limit the scope of patient demographic data accessed in the manual review component of the HIO patient lookup process. Appropriate privacy and security policies and processes should be established to ensure that patient privacy is not placed at unreasonable risk in the pursuit of efficiencies. To limit inadvertent disclosure of clinical data, false positive rates should be well understood and minimized. The potential volume of demographic data that an HIO shares when users attempt to disambiguate inexact matches may require further exploration.

Conclusion 3. Research into methods of patient matching can benefit from a review of systems used in other business sectors. Patient matching is a specific instantiation of a process that falls under the more general rubric of *identity resolution*, which refers to the process of discovering linkages and obscure relationships within and among data sources. In addition to health care–specific patient matching, identity resolution is implemented across many business sectors, including law enforcement, financial services (including banking and insurance), government, and retail services. Financial services and law enforcement benefit from identity resolution by identifying potentially dubious connections between persons using demographic data, geographical locations, financial transactions, and other data elements to discover fraudulent and criminal activity. Nationwide credit scores are tabulated by integrating (linking) person-based data across many sources. In the retail and other business sectors, customer relationship management (CRM) systems incorporate identity resolution methods to ensure that businesses maintain up-to-date information on customers and clients. Many common government functions, such as motor vehicle services and voter registration boards, depend on high-quality lists. Identity resolution methods are used to maintain these lists, and solutions for many patient matching challenges may come from innovations implemented in other business sectors.

An important consideration in using innovations from other sectors is that business sectors outside of health care may fall under less restrictive privacy and security laws and regulations and therefore may be able to combine many different sources of data without the patient’s/person’s permission, which is often required in health care situations. Although health care can and does benefit from the work in other sectors, the application of these innovations may be tempered by existing laws and policies on the use and disclosure of health information.

Conclusion 4. The health care industry should continue to share best practices. Creating accurate and computationally feasible matching strategies is a complex challenge. By sharing best practices that are supported by formal evaluation of real-world matching challenges, the barriers to creating accurate matching strategies for electronic health information exchange can be lowered. This change will lead to improved matching processes, which in turn will provide more accurate and complete clinical health information, both at the point of health care delivery and for clinical research.

REFERENCES

American National Standards Institute. (2008). *HITSP patient ID cross-referencing transaction package* (HITSP/TP22). Retrieved June 22, 2009, from http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=2&PrefixNumeric=22

Appavu, S. I. (1997). *Analysis of unique patient identifier options—Final report.* Washington, DC: U.S. Department of Health and Human Services.

Appavu, S. I. (1999). Unique patient identifiers: What are the options? *Journal of AHIMA, 78*(3), 34–37.

Baxter, R., Christen, P., & Churches, T. (2003). A comparison of fast blocking methods for record linkage. In *Proceedings of the ACM SIGKDD'03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation* (pp. 25–27).

Bondi, A. B. (2000). Characteristics of scalability and their impact on performance. In *Proceedings of the 2nd International Workshop on Software and Performance* (pp. 195–203). Ottawa, Ontario, Canada.

Campbell, K. M., Deck, D., & Krupski, A. (2008). Record linkage software in the public domain: A comparison of Link Plus, the Link King, and a "basic" deterministic algorithm. *Health Informatics Journal, 14*(1), 5–15.

Christen, P. (2006). A comparison of personal name matching: Techniques and practical issues. In *Proceedings of the Workshop on Mining Complex Data (MCD) held at the IEEE International Conference on Data Mining.* Hong Kong, China.

Christen, P. (2008). *Febrl: A freely available record linkage system with a graphical user interface.* Paper presented at the Australasian Workshop on Health Data and Knowledge Management, Wollongong, New South Wales, Australia.

Christen, P., & Pudjijono, A. (2009). *Accurate synthetic generation of realistic personal information.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand.

Cohen, W. W., & Richman, J. (2002). Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 475–480). Edmonton, Alberta, Canada.

Dal Maso, L., Braga, C., & Francheschi, S. (2001). Methodology used for "software for automated linkage in Italy" (SALI). *Journal of Biomedical Informatics, 34*(6), 387–395.

Dimitropoulos, L. (2007a). *Privacy and security solutions for interoperable health information exchange: Assessment of variation and analysis of solutions.* Rockville, MD: Agency for Healthcare Research and Quality.

Dimitropoulos, L. (2007b). *Privacy and security solutions for interoperable health information exchange: Final implementation plans.* Rockville, MD: Agency for Healthcare Research and Quality.

Dimitropoulos, L. (2007c). *Privacy and security solutions for interoperable health information exchange: Nationwide summary*. Rockville, MD: Agency for Healthcare Research and Quality.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16.

Fellegi, I. P., & Sunter, S. B. (1969). A theory of record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210.

Fernandes, L., & O'Connor, M. (2008). Patient identification in three acts. *Journal of AHIMA, 79*(4), 46–49.

Grannis, S. J., Overhage, J. M., & McDonald, C. J. (2002). Analysis of identifier performance using a deterministic linkage algorithm. In I. S. Kohane (Ed.), *Proceedings of the AMIA 2002 Annual Symposium* (pp. 305–309). Bethesda, MD: American Medical Informatics Association.

Grannis, S. J., Overhage, J. M., & McDonald, C. J. (2003). *Analysis of a probabilistic record linkage technique without human review.* Paper presented at the American Medical Informatics Association Fall Symposium, Washington, DC.

Greenberg, M. D., & Ridgely, M. S. (2008). Patient identifiers and the National Health Information Network: Debunking a false front in the privacy wars. *Journal of Health and Biomedical Law, 4*(1), 31–68.

Gu, L., Baxter, R., Vickers D., & Rainsford, C. (2003). *Record linkage: Current practice and future directions* (Technical Report No. 03/83). Sydney, New South Wales, Australia: CSIRO Mathematical and Information Sciences, CMIS.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques.* New York: Springer.

Hillestadt, R., Bigelow, J. H., Chaudhry, B., Dreyer, P., Greenberg, M. D., Meili, R. C., Ridgely, M. S., Rothenberg, J., & Taylor, R. (2008). *Identity crisis: An examination of the costs and benefits of a unique patient identifier for the U.S. health care system* (MG-753-HLTH). Santa Monica, CA: RAND Corporation.

Jaro, M. (1999). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. In *Record Linkage Techniques—1997: Proceedings of an international workshop and exposition* (pp. 351–357). Arlington, VA: National Academy Press.

Kansky, J. (2007, August 2). Value, relevance key to HIE success. *Healthcare IT News*. Retrieved June 22, 2009, from http://www.healthcareitnews.com/news/value-relevance-key-hie-success

Karmel, R., & Gibson, D. (2007). Event-based record linkage in health and aged care services data: A methodological innovation. *BMC Health Services Research, 7*(154). Retrieved June 22, 2009, from http://www.biomedcentral.com/1472-6963/7/154

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Liu, S., & Wen, S. W. (1999). Development of record linkage of hospital discharge data for the study of neonatal readmission. *Chronic Diseases in Canada, 20*(2), 77–81.

Lyons, R. A., Jones, K. H., John, G., Brooks, C. J., Verplancke, J.-P., Ford, D. V., Brown, G., & Leake, K. (2009). The SAIL databank: Linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making, 9*(3). Retrieved June 22, 2009, from http://www.biomedcentral.com/1472-6947/9/3

Markle Foundation. (2005). *Linking health care information: Proposed methods for improving care and protecting privacy.* New York: Author. Retrieved June 22, 2009, from https://www.policyarchive.org/handle/10207/15521

Markle Foundation. (2006a). *Connecting for Health Common Framework: Correctly matching patients with their records.* New York: Author. Retrieved June 22, 2009, from http://www.connectingforhealth.org/commonframework/docs/P4_Correctly_Matching.pdf

Markle Foundation. (2006b). *Connecting for Health Common Framework: Record Locator Service: Technical background from the Massachusetts Prototype Community.* New York: Author. Retrieved June 22, 2009, from http://www.connectingforhealth.org/commonframework/docs/T6_RecordLocator.pdf

Meray, N., Reitsma, J. B., Ravelli, A. C., & Bonsel, G. J. (2007). Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *Journal of Clinical Epidemiology, 60*(9), 883–891.

Mohamed, G., Elfeky, G., Verykios, V., & Elmagarmid, A. (2002). TAILOR: A record linkage toolbox. In *Proceedings of the 18th International Conference on Data Engineering* (p. 17). Washington, DC: IEEE Computer Society.

Monge, A. E. (2000). Matching algorithms within a duplicate detection system. *IEEE Data Engineering Bulletin, 23*(4), 14–20.

Morrissey, J. (2007). *Safety in numbers: Resolving shortcomings in the matching of patients with their electronic records.* Chicago: National Alliance for Health Information Technology.

National Alliance for Health Information Technology. (2008). *Defining key health information technology terms.* Washington, DC: Office of the National Coordinator for Health Information Technology. Retrieved June 22, 2009, from http://www.nahit.org/images/pdfs/NAHIT_Key_HIT_Terms_Report.pdf

National E-Health Transition Authority. (2006). *Privacy blueprint—Unique healthcare identifiers: Individual healthcare identifiers and healthcare provider identifiers, Version 1.0.* Sydney: National H-Health Transition Authority, Ltd.

Netter, W. (2003). Curing the unique health identifier: Reconciliation of new technology and privacy rights. *Jurimetrics, 43*(2), 165–186.

Newman, T., & Brown, A. (1997). Use of commercial record linkage software and vital statistics to identify patient deaths. *Journal of the American Medical Informatics Association, 4*(3), 233–237.

Pates, R. D., Scully, K. W., Einbinder, J. S., Merkel, R. L., Stukenborg, G. J., Spraggins, T. A., Reynolds, C., Hyman, R., & Dembling, B. P. (2001). Adding value to clinical data by linkage to a public death registry. *MedInfo, 10*(2), 1384–1388.

Porter, E. H., & Winkler, W. E. (1999). Approximate string comparison and its effect on an advanced record linkage system. In *Record Linkage Techniques—1997: Proceedings of an international workshop and exposition* (pp. 190–202). Arlington, VA: National Academy Press.

Prabhakar, S., Pankanti, S., & Jain, A. K. (2003). Biometric recognition: Security and privacy concerns. *IEEE Security & Privacy, 1*(2), 33–42.

Rollins, G. (2007). This year's models: A look at patient ID in the four newly demonstrated NHIN prototypes. *Journal of AHIMA, 78*(3), 34–37.

Sauleau, E. A., Paumier, J. P., & Buemi, A. (2005). Medical record linkage in health information systems by approximate string matching and clustering. *BMC Medical Informatics and Decision Making, 5*(32). Retrieved June 22, 2009, from http://www.biomedcentral.com/1472-6947/5/32

Shekelle, P. G., Morton, S. C., & Keeler, E. B. (2006). *Costs and benefits of health information technology* (AHRQ Publication No. 06-E006). Rockville, MD: Agency for Healthcare Research and Quality.

Sideli, R. V., & Friedman, C. (1991). Validating patient names in an integrated clinical information system. In *Proceedings of the Annual Symposium on Computer Applications in Medical Care* (pp. 588–592). Washington, DC: American Medical Informatics Association.

Sloane, E. B., & Carey, C. C. (2007). Using standards to automate electronic health records (EHRs) and to create integrated healthcare enterprises. *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6178–6179).

Social Security Administration. (2001). *Unresolved death alerts over 120 days old* (Audit Report A-09-00-10001). Washington, DC: Office of the Inspector General.

Stewart, S. P., Arellano, M. B., & Simborg, D. W. (1984). Optimal patient identification system. *Journal of the American Medical Record Association, 55*(8), 23–27.

Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D. W., & Middleton, B. (2005, January 19). The value of health care information exchange and interoperability. *Health Affairs* (Web Suppl.). Retrieved June 22, 2009, from http://content.healthaffairs.org/cgi/content/full/hlthaff.w5.10/DC1

Whalen, D., Pepitone, A., Graver, L., & Busch, J. (2001). *Linking client records from substance abuse, mental health and Medicaid state agencies* (SAMHSA Publication No. SMA-01-3500). Rockville, MD: Substance Abuse and Mental Health Services Administration.

Winkler, W. E. (2000). *Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage* (Statistical Research Report Series No. RR2000/05). Washington, DC: Statistical Research Division, U.S. Bureau of the Census.

Winkler, W. E. (2006). *Overview of record linkage and current research directions* (Statistics Research Report Series No. 2006-2). Washington, DC: Statistical Research Division, U.S. Census Bureau. Retrieved June 22, 2009, from http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

Wooster, L. (2006). *Identification of patients: A technical and cultural challenge that can no longer be avoided.* Chicago: National Alliance for Health Information Technology.

APPENDIX A
HIO INTERVIEW QUESTIONS

1. [So that we can] better understand the size of your health information exchange (HIE), please answer the following questions about your HIE patient population:

    a. How many unique patients are recorded in your HIE?
    b. How many patient registrations are recorded in your HIE (this likely includes multiple registration events per unique patient)?
    c. How many distinct registration domains participate in your HIE? (A registration domain represents an entity that independently registers and tracks patient information for a variety of health care purposes; ideally, each patient has no more than one identity within a registration domain. Examples of registration domains include hospitals, free-standing laboratories, and outpatient clinics. Alternatively, the three previous examples may be present under the same registration domain if they share the same enterprise master person index.)

2. What software components do you use to perform the data matching/linkage?
    a. Did you choose a commercial off-the-shelf solution? If so, which vendor?
    b. Did you develop and implement your own matching solution?
    c. Do you use a combination of commercial and internally developed solutions?

3. Would you characterize your linkage process as probabilistic (statistical) or deterministic (rule-based)?

4. Could you please describe the general architecture of your HIE patient matching method (distributed or centralized)? For example, is the linkage process implemented in a distributed, peer-to-peer fashion? Or alternatively, is a copy of identifying patient data maintained in a centralized location, and centralized matching process?

5. Do you manually review questionable matches? If so, approximately how many full-time equivalent (FTEs) [employees] are allocated to adjudicating matches?

6. How do you address patient security and confidentiality as it pertains to patient matching?

    a. When searching for a patient, does your matching program allow you to view multiple possible match candidates? Are candidates constrained in any way?
    b. Does your matching program return more information (identifying fields) than originally entered to find the patient? Are any fields obfuscated?
    c. How is access to the matching function controlled (who can search for patients)?

7. Are a minimum number of variables required to return any match candidates?

8. Have you established an acceptable level for false positive matches?

    a. If so, how was it established?
    b. Do you have processes in place to identify and adjudicate potential false positives?
    c. Do you vary the acceptable level of false positives for different use-cases (for newborn screening matching, for example)?

9. How do you identify and adjudicate potential twins or familial links?

10. Do you have a mechanism to incorporate feedback from the registration domains involved in the HIE? For example, what do you do if an entity informs you of an incorrect match or patient identity theft?
11. If a potential match is not found during the matching process, does the matching software inform the user?
12. If there is a match but the results are restricted in some—either due to sensitive data or because the provider does not have permission to view the data—does the matching software inform the user? And if so, what message does the program display?