Department of Health & Human Services
Office of the National Coordinator for Health IT
HIT Policy Committee
Privacy & Security Tiger Team
Patient Linking Hearing
Thursday, December 9, 2010

Written Testimony

Shaun J. Grannis, MD MS FAAFP
Medical Informatics Research Scientist, The Regenstrief Institute
Director, Indiana Center of Excellence in Public Health Informatics
Assistant Professor, Dept of Family Medicine
Indiana University School of Medicine

Introduction

I want to express my gratitude to the Privacy and Security Tiger Team members for the opportunity to engage in a discussion on patient linking, a topic of vital importance to nationwide health information exchange.

First and foremost, the testimony presented is based on both the practical experience of developing and maintaining a global patient identity resolution service for the Indiana Network for Patient Care (INPC). Accurately aggregated clinical data is a fundamental component of the INPC, and enables a variety of the use cases illustrated in figure 1. With data extending back over 30 years, the INPC is one of the nation's most comprehensive and longest tenured health information exchanges, containing more than 3.1 billion coded standardized clinical observations, and a global patient index containing more than 20 million person:source entities that represent more than 12 million unique persons. Since the mid-1990's, the INPC global patient identity resolution service resolves identities from real-time clinical data streams provided by myriad sources with widely varying data quality. Currently, the INPC global patient identity resolution service adjudicates identities for between 350,000 and one million transactions received from over 1,100 distinct participating HIE sources.

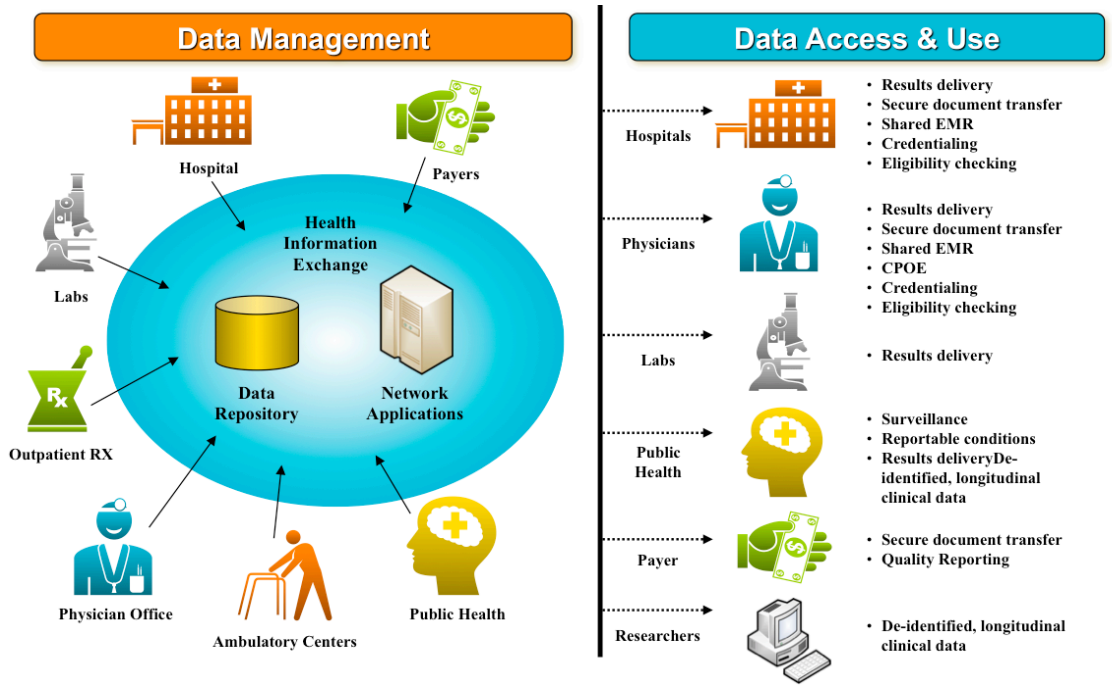

Figure 1: Data infrastructure and use cases supported by the Indiana Network for Patient Care (INPC).

Second, the testimony presented is informed by ongoing research evaluating various dimensions of the patient linkage process.[1-4] Those dimensions include evaluating data quality, developing novel statistical methods, comparing the accuracy of various linkage methods, and assessing the impact of clinical context on approaches to linkage methodologies.

Third, the testimony presented reflects insights derived from serving in a consultative capacity [5-6], including serving as the lead technical author for the White Paper entitled, "Perspectives on Patient Matching: Approaches, Findings, and Challenges," [5] prepared for the Office of the National Coordinator for Health IT (ONC) and the Agency for Healthcare Research and Quality (AHRQ), published June 30, 2009.

What is the problem?

Health care information is distributed across many independent databases and systems, both within and among organizations as separate islands with different patient identifiers. This is the case for data collected within an institution where there may be multiple identifiers, or for data collected about the same patient at different health care institutions, different pharmacy systems, different payers, and so on. This situation interferes with the aggregation of information about individuals as needed for processes such as public health reporting, clinical research, outcomes management, and quality reporting. Aggregation is important not only to determine a patient's health care status, but also for population-based studies. *Patient linkage* is the process of combining information about an individual residing in one or more databases.

Contributing to the linkage challenge is the reality that identifiers for a given individual are variably recorded for a variety of reasons listed in table 1. And while unique identifiers such as the Social Security Number (SSN) or a national healthcare identifier [7-9] may be appealing, they are not a panacea for a variety of reasons. Table 2 highlights scenarios derived from actual data where false-positive links would occur when SSN is used as the only linking variable. Further, if a national healthcare identifier strategy is sought, a return on investment analysis must incorporate the estimated cost for deployment of between $4.9 billion and $12.2 billion (adjusted to 2009 dollars).[7-8]

Table 1: Common causes for identifier variation

- Recording Errors
  - Phonetic ("Shaun", "Sean", "Shawn")
  - Typographical
    (Smith → Snith, "07" → "01")
- Changing Identifiers
  - Last Name (Marriage)
  - Geographic location (Home address, etc)
- Sharing Identifiers (SSN, etc.)
- Identifiers Limited or Unavailable

Processes and standards for identifying individuals

The choice of patient linking approach is strongly influenced by a variety of factors including the need for human supervision, workflow timing constraints, and desired linkage accuracy. 'Human supervision' refers to whether manual resources are available for disambiguating uncertain matches, or whether aggregation will be performed in a fully automated fashion. Workflow timing refers to whether aggregation 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.

Table 2: Three types of false positives from SSN linkage [1]. The most common error appear to be due to familial mix-ups in that a female of one record was linked to a male record sharing the same last name. Typographical errors and SSN collisions of unknown etiology accounted for the remainder of the errors. Examples use fictitious data.

	SSN	LN	FN	G	MB	DB	YB
Typographical Errors							
Institution Data	123446789	SMITH	FRED	M	12	20	1908
SSDMF Correct Name Link	123456789	SMITH	FRED	M	12	20	1908
SSDMF Incorrect SSN Link	123446789	JONES	PAT		7	13	1914
Spousal Linkage							
Institution Data	987654321	COLLINS	MARY	F	8	29	1917
SSDMF Correct Name Link	987708345	COLLINS	MARY	F	8	29	1917
SSDMF Incorrect SSN Link	987654321	COLLINS	TIM	M	4	3	1915
Unexplained Collisions							
Institution Data	586192376	PRATT	DAVID	M	10	3	1932
SSDMF Correct Name Link	666248764	PRATT	DAVID	M	10	3	1932
SSDMF Incorrect SSN Link	586192376	WILSON	ERIN		5	17	1940

Strategies for data aggregation use-cases fall into one of four categories based on the constraints of human supervision and timing (Figure 2). Batch mode matching with human supervision is often used for research and creating aggregate reports where high accuracy is required. HIE's may not be able to afford the high cost of a human operator because human review can require thousands of man-hours of work, or because privacy concerns dominate. In such a case, HIE's may choose to automatically link patient data

in real-time without human supervision. De-identified (anonymous) 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. Finally, an example of real-time matching with human supervision is implemented in most large hospital systems where many thousands of patient transactions are aggregated daily, and humans disambiguate uncertain matches. Human resources are devoted to disambiguation in this case because hospitals derive value from accurate patient data, which is used for revenue generating business processes such as billing and invoicing. These distinctions are important, because much published patient linkage research has done conducted researchers with modest size data sets, as a one time batch effort, which is different from HIE's where matching is done on the context of an ongoing, dynamic and large system.

Human Supervision	Workflow Timing: Batch-mode	Workflow Timing: Real-time
Substantial Manual Supervision	**Reporting, Research**	**Health Care Enterprise (Hospitals)**
Little or no Manual Supervision	**De-identified Matching**	**Health Information Exchange**

Figure 2: Examples of patient linking scenarios broken divided by dimensions of workflow timing and human supervision.

INPC Matching Process

For patient care purposes, the INPC global patient index contains records containing patient demographics collected from many registration domains, but no clinical data. These records act as pointers to the specific, multiple federated clinical repository containing patients' actual clinical data. The global patient registry is centralized while the federated clinical repositories can be held separately, or in the same location.

Each patient has at least one entry per registration domain. Because patients often receive care at different health care organizations, they often have more than one entry in the global patient index. A collection of registration records for the same patient is

called a 'patient group'. The INPC receives over 350,000 clinical transactions per day in real time, and all transactions must be linked to a patient. If the transaction contains unchanged, previously linked demographic and medical record number (MRN) data from the given domain, it is automatically linked to the existing patient group. If demographic data has changed, the transaction is flagged for linking using the full linkage algorithm. The linkage process can analyze flagged records multiple times per day. To access clinical data, clinical applications query the global patient identity resolution service for all locations (registration domain) of a given patient's data. Once all locations are identified, the appropriate clinical data is retrieved from the federated clinical data repositories. Because we implement a highly specific algorithm (low false-positives) that favors a false non-link rather than a false-link, fiercely defending against the false link.

While the INPC linkage algorithm is tuned for high specificity and positive predictive value, the desired linkage performance characteristics may vary for different scenarios. Consequently, it's important to recognize the trade-off between false positives and false negatives.

Although rejecting all false positives may be desirable, it is impractical because false positive and false negatives are inversely related. In general, as the false positive rate is reduced, the false negative rate increases (see Figure 3). 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.

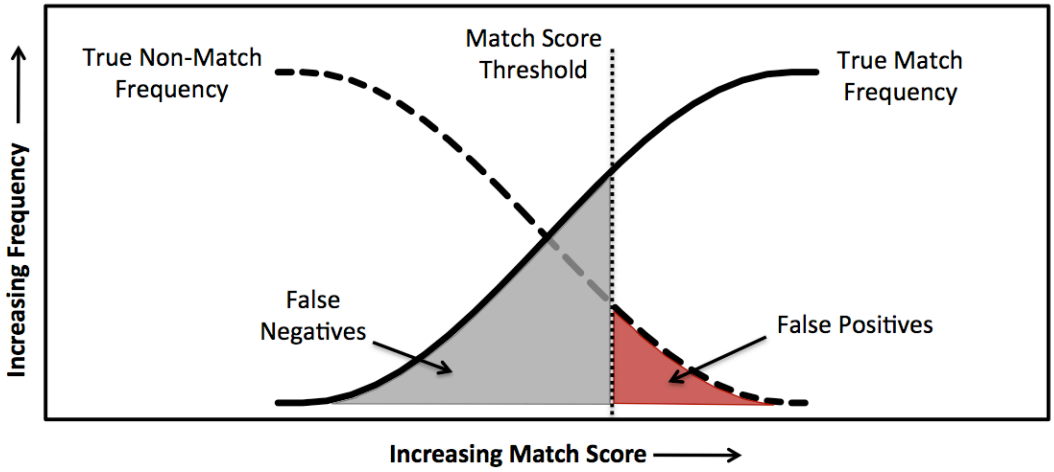

Figure 3: The relationship between false positive and false negative matches. As the match score threshold is increased, the number of false positives decreases, but false negatives increase. As the match score threshold is lowered, the number of false negatives decreases, but false positives increase.

Other INPC processes may require differing patient linkage performance characteristics. One such process involves improving newborn screening follow-up by identifying infants who may lack screening. Not all infants are appropriately screened for harmful or potentially fatal disorders that are otherwise unapparent at birth. Although public health authorities can link vital records data with newborn screening results to identify unscreened infants, such processes may be delayed and some cases may remain undetected by this process. To improve detection of unscreened infants, we have developed a process to verify that all infants within our HIE have evidence of a newborn screen. To accomplish this, we can monitor real-time data streams and attempt to match each infant within the HIE to Indiana's statewide newborn screening registry using a robust matching process. In addition to the typical challenges of missing SSN's and first names in newborns, matching infants in an HIE is particularly challenging. This is because manually gathering health care data is costly and health care transactions typically contain the minimal amount of data to accomplish the task for which the transaction was intended. Further, while the public health newborn registry offers a consistent set of matching fields collected using a standard form, the data fields collected across various healthcare clinics, labs and ancillary services organizations within an HIE record a variety of different fields. For example, systems inconsistently gather next-of-kin and provider information, which can improve linkage accuracy when present.

To achieve the desired linkage accuracy for this relatively small cohort with substantial data quality issues, we use a probabilistic algorithm that very flexibly accommodates a broad range of informational characteristics.

Understanding the Costs

In general, if false positive links are more costly, we can configure the probabilistic linkage algorithm to function at higher specificity (or positive predictive value); if false negative links are more costly, we can configure the algorithm to function at higher sensitivity. The perceived costs of incorrect patient linkages influences the evaluation of linkage algorithms, and perceived costs can be influenced by clinical context and intended use of the linked data. For example, when facing an ambiguous potential patient link (see below), should these two records be linked if it was known that at least one of the entities was diagnosed with HIV? Alternatively, is it appropriate to link these two records as part of an ultimately de-identified, aggregate population-level summary?

SSN	LAST	FIRST	MID	YB	DB	MB	SEX	TEL	STREET	CITY	ST	ZIP
303889999	THOMOSANI	ROBERT	L	1985	7	20		000-555-6666	302 ELM ST	INDIANAPOLIS	IN	46278
303889999	THOMPSON	ROBERT		1955	7	20	M	317-555-6666	302 ELM ST	SPEEDWAY	IN	46277

Evaluating linkage output to ensuring accurate patient linkage

Patient linkage methods are typically evaluated using one of two approaches. First, the output of a data aggregation 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 full manual review of all linkages is usually infeasible due to the large volume of records processed, manual review is typically performed on a fractional sampling of the data, and statistical inferences are made. Second, synthetic data that contains known linkage outcomes can be used to create an a-priori gold-standard against which the algorithm's performance can be assessed. While this approach is attractive because it may obviate the need for more costly manual review, validity of synthetic evaluations are strongly dependent on the degree to which the synthetic data reflects the underlying characteristics of the targeted real-world data.

Ideally, evaluations of patient linkage approaches in the context of large data volumes should (a) be based on a "gold standard" set of true links that is reflective of the entire data set being aggregated, and (b) ideally be assessed using identical data sets. However, (a) due to the typically large number of potential links generated by record linkage processes (not uncommonly hundreds of millions of records), it is infeasible to establish a "gold standard" through manual review of real data. Further, item (b) implies that the data to be linked should be accessible to different algorithm developers if cross-algorithm comparison is desired. However, sharing linkage datasets containing actual patient data is largely prohibited due to privacy and confidentiality restrictions governing identifiable patient data. Thus, patient linkage system stakeholders and evaluators face barriers to assessing algorithms and sharing data.

Synthetic data may aid in future algorithm evaluation. Processes that create synthetic data can be used to construct de-identified a-priori "gold-standards" against which an algorithm's performance can be more easily measured. However, as previously stated the validity of synthetic data-based evaluations are strongly dependent on the degree to which the synthetic data reflects the underlying characteristics of the targeted real-world data. Therefore, to maximize the validity of such evaluations, synthetic data should closely reflect the underlying characteristics of the original data, while avoiding re-identification risks. This is a crucial issue because moderately varying data characteristics can strongly influence algorithm performance. There is a need for work in this space because existing efforts have focused on creating rule-based models for specific field errors rather than building a non-identifying, parameterized statistical model that reflects characteristics of the underlying data.

Given that real-world healthcare data is imperfect and given the inherent trade-off between false positives and false negatives, we optimize patient linkage for patient care purposes in the INPC through a number of mechanisms. First, we begin by designing the algorithm with the aim of higher specificity and positive predictive value rather than higher recall or sensitivity. Second, the INPC global patient identity resolution service retains distinct entities for each patient registration domain and does not merge patient identities permitting seamless unlinking of patients and their respective data. Consequently, if an INPC participant identifies an incorrect link, it can be manually corrected. Third, we can routinely examine suspect links, e.g., links with disagreeing

gender, for false positives and we can use a “greedy” (high sensitivity/recall) probabilistic algorithm to manually explore missed links.

Barriers to accurate patient-matching: Data Quality

While our linkage analyses indicate high precision (positive predictive value) and recall (sensitivity), these same analyses suggest that improved data quality (e.g., improved sources-system recording accuracy and completeness) would further enhance linkage performance. Both experience and an apocryphal maxim suggest:

> “*When given the choice between good data and a good linkage algorithm, choose good data.”*

However, improving data quality often requires in-depth workflow evaluation and system re-design, processes that are resource-intensive and whose return on investment is oft not easily or immediately realized by an organization. Additionally, transactions for specific cohorts such as newborns often lack discriminating identifiers such as social security number and consistent first name because these identifiers legitimately do not exist at the time of the transaction. Further, patients resulting from multiple gestation births (twins, triplets, etc.) can share highly similar identifying information, leading to potential false links. Birth order information, which aids in preventing false links among siblings, is often missing from routine healthcare transactions.

Consequences of a wrong match

We are unaware of any false links in our system leading to inappropriate and/or harmful care, nor are we aware of any inadvertent breaches in privacy or confidentiality as a result of inappropriate links. While the most likely consequence of a false positive is the inadvertent disclosure of protected health information, the potential for adverse clinical outcomes is less certain [5]:

> *The chances of a false positive causing serious harm are rare—it requires that the false positive occur and that the incorrect information be [fully consistent with the presenting clinical scenario]. For example, [a] contraindication for thrombolytic therapy is recent heart surgery. However, if the clinical record indicates 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, HIO's 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.

What should ONC do to address patient matching problems in information exchange?

The following recommendations are taken from [5], page 8-1 through 8-4. The recommendations are divided into "Functional Aspects and Transparency" and "Conceptual and Policy Related":

Functional Aspects and Transparency

(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.

(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, 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. 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.*

(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.

Conceptual and Policy-Related Matters

(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.

(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.

(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. 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 when assessing innovations from other sectors is that non-health care sectors 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, existing laws and policies governing the use and disclosure of health information may temper the application of these innovations.

(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.

References:

1. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. Proc AMIA Symp. 2002:305-9. PubMed PMID: 12463836.

2. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. AMIA Annu Symp Proc. 2003:259-63. PubMed PMID: 14728174

3. Grannis SJ, Overhage JM, McDonald C. Real world performance of approximate string comparators for use in patient matching. Stud Health Technol Inform. 2004;107 (Pt 1):43-7. PubMed PMID: 15360771.

4. Zhu VJ, Overhage MJ, Egg J, Downs SM, Grannis SJ. An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):738-45. PubMed PMID: 19567789

5. Grannis SJ, Banger A, Harris D. (2009). Perspectives on Patient Matching: Approaches, Findings, and Challenges [White paper]. Office of the National Coordinator for Health Information Technology. Retrieved from http://healthit.hhs.gov/portal/ server.pt?open=512 &objID=1240&mode=2.

6. UNAIDS/PEPFAR. (2009) Interim Guidelines for the Development and Use of Individual Identifiers for the Provision of Health Services including HIV: Proceedings from a Workshop, 24–26 February 2009, Montreaux, Switzerland [White paper].

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

8. 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.

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