Final Progress Report

Project Title:

Advancing Patient Identity Management in the Context of Real-World Health Information Exchange

Principal Investigator:

Shaun Grannis, MD, MS, FAAFP

Team Members:

Siu Hui
Huiping Xu
Judy Gichoya
Joanne Daggy
James Egg
Jane Wang
Andrew Martin
Emily Mitchell
Jessica Craig
Joseph Gibson

Organization:

The Trustees of Indiana University

Project Dates:

September 30, 2009 – July 31, 2013

Federal Project Officer:

Denise Burgess

Acknowledgement of Agency Support:

This project was supported by grant number R01HS018553 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Grant Number:

5R01HS018553

Inclusion of AHRQ Priority Populations

- Inclusion of Children
    - Because we collected data directly from source systems, we did not know which category the patient might belong to. We did not specifically target children for inclusion in this study; however, their involvement was anticipated, as we examined data regardless of patient category. Please note that this study involved no direct interaction with patients, and the risk of loss of privacy was minimal, as outlined in the human subjects section.
    - Targeted and Planned Enrollment
        - We planned to potentially include all patients for whom data are captured in the INPC health information exchange. There are currently data for over 12 million unique patients with over 25 million associated registration events stored in the system. These patients reflect the population of the state of Indiana broadly in terms of age distribution, gender, and race.
- Inclusion of women and minorities
    - Because we collected data directly from source systems, we cannot know which category the patient might belong to. All women and minorities who have or will present to a participating INPC institution for healthcare treatment will be included in this study. Please note that this study involved no direct interaction with the patients, and the risk of loss of privacy is minimal, as outlined in the human subjects section.
- Protection of Human Subjects
    - Risks to Human Subjects
        - Human Subjects Involvement and Characteristics
            - The proposed research relied exclusively on existing electronic medical records. The patients were not contacted or intervened on in any way.
            - The data contained in this system represented patients from across the state of Indiana. Specifically, we looked at sections of the record that allow us to match records across source systems. We did not analyze the reasons for or outcomes of patient visits.

<u>Proposed</u> Enrollment Table

(as submitted in original proposal)

Study Title: Advancing Patient Identity Management in the Context of Real-World Health Information Exchange
Total Estimated Enrollment: 10,000,000

PROPOSED ENROLLMENT: Number of Subjects by Registration				
Ethnic Category	**Sex/Gender**			
	Females	**Males**	**Unknown or Not Reported**	**Total**
Hispanic or Latino	178,500	171,500		350,000
Not Hispanic or Latino	4,921,500	4,728,500		9,650,000
Unknown (individuals not reporting ethnicity)				
Ethnic Category: Total of All Subjects	5,100,000	4,900,000		10,000,000
Racial Categories				
American Indian/Alaska Native	51,000	49,000		100,000
Asian	0	0		0
Native Hawaiian or Other Pacific Islander	15,300	14,700		30,000
Black or African American	443,700	426,300		870,000
White	4,590,000	4,410,000		9,000,000
Unknown or Not Reported				
Racial Categories: Total of All Subjects	5,100,000	4,900,000		100,000

Actual Enrollment Table

Study Title: Advancing Patient Identity Management in the Context of Real-World Health Information Exchange
Total Actual Enrollment: 24,126,855

ACTUAL ENROLLMENT: Number of Subjects by Registration				
Ethnic Category	**Sex/Gender**			
	Females	**Males**	**Unknown or Not Reported**	**Total**
Hispanic or Latino				
Not Hispanic or Latino				
Unknown (individuals not reporting ethnicity)	12,237,142	9,366,736	2,522,977	24,126,855
Ethnic Category: Total of All Subjects	12,237,142	9,366,736	2,522,977	24,126,855
Racial Categories				
American Indian/Alaska Native				
Asian				
Native Hawaiian or Other Pacific Islander				
Black or African American				
White				
Unknown or Not Reported	12,237,142	9,366,736	2,522,977	24,126,855
Racial Categories: Total of All Subjects	12,237,142	9,366,736	2,522,977	24,126,855

Structured Abstract

Purpose

We evaluated patient matching processes in an operational HIE, applied novel modifications to existing matching algorithms, and assessed the value of matching across clinical sources both within and outside the HIE.

Scope

We used data from the Indiana Network for Patient Care, the nation's largest and longest-tenured HIE. We performed *linkage* analyses across various clinical data sources in the HIE and sources external to the HIE; we also conducted *deduplication* analyses among individual data sources by matching data sources to themselves.

Methods

We implemented and evaluated extensions to probabilistic methods designed to improve accuracy. Extensions included stochastic and closed-form approaches to parameter estimation; generalization of probabilistic methods to accommodate dependence between fields; and evaluation of methods detecting the presence of specific data characteristics that inform the selection fields used by the matching model. We assessed processes for identifying data element combinations failing the test for independence. We evaluated and characterized the accuracy of matching real-world HIE data sources for a variety of scenarios using both deterministic and probabilistic methods.

Results

Through various analyses, we demonstrated that, although matching methods can be feasibly applied to a variety of data, barriers to optimal matching remain, and there is a paucity of evidence-based best practices for matching approaches. Our work provides evidence-based guidance for accommodating model variations, such as field dependence; we described methods for efficiently sampling records to establish "gold-standard" reference datasets used for measuring accuracy.

Key Words
Record linkage; deduplication; data quality; health information exchange; patient identification

Purpose

Comprehensive clinical information from a broad set of data sources is required for many healthcare tasks, including comparative effectiveness research and improving the value of care delivered. Health information exchanges are an emerging source of comprehensive clinical data for these purposes. However, HIE data are captured from many independent databases and systems in which data are stored as separate islands with different patient identifiers. This impedes the aggregation of information about individuals across such data sources as needed for many uses, including routine delivery of care, public health processes, clinical and comparative effectiveness research, and other healthcare-related processes. Therefore, accurate and robust record linkage capabilities are needed. Because HIEs are an amalgamation of many data sources, with data quality characteristics that vary both by data source and by time, they pose particularly difficult record linkage challenges. Though data aggregation and record linkage methods have been studied previously, there have been no formal, comprehensive evaluations of record linkage methodologies using real-world heterogeneous healthcare transactions that reflect health information exchange data characteristics. Linkage evaluations to date have not reflected the challenges faced by HIEs for the following reasons: First, analyses often used data derived from small numbers of sources, sources with similar data characteristics, or similar workflows. Second, these data often undergo extensive pre-processing or cleaning prior to linking. These factors produce consistent, well-known data characteristics. Third, some evaluations are misleading, because they reflect the success of the algorithm assisted by human review, not the algorithm alone. The amount of human review required to achieve the level of accuracy described in published analyses may be infeasible for HIEs, because they can receive hundreds of thousands of transactions daily from many sources with varying data characteristics. Thus, evaluations of the performance of robust, accurate, unsupervised linkage methods in many clinical contexts reflected by HIEs are needed. Consequently, we sought to:

1. Evaluate various classes of linkage methodologies currently used in the context of a long-standing, operational health information exchange: (1) a deterministic fuzzy-match linkage algorithm used for unsupervised, real-time linkage, and (2) a probabilistic linkage algorithm with novel modifications also used for unsupervised linkage.

2. Implement and evaluate extensions to the probabilistic linkage method that are designed to improve algorithm accuracy by overcoming shortcomings in existing probabilistic linkage models. Extensions include stochastic and closed-form solutions for parameter estimation methods; generalization of the probabilistic method is used to accommodate statistical dependence between fields. We characterized the relative performance gains using appropriate metrics such as sensitivity (recall), positive predictive value (precision), AIC, BIC, and AUC. To characterize the performance of algorithm extensions, we used a variety of evaluation approaches, including manually reviewing actual linked data.

3. Investigate methods that detect the presence or absence of specific data fields and data characteristics that inform the selection of extensions to the underlying probabilistic matching model. We developed and evaluated processes for identifying data element combinations that fail the test for statistical independence.

4. Evaluate and characterize technical performance and clinical and operational value of linking real-world HIE data sources, comparing the effectiveness of both deterministic and probabilistic methods. Clinical and public health scenarios included evaluating the completeness and efficiency of aggregating clinical data at the point of care in an operational HIE; identifying duplicate patients and missing linkages within and between a variety of special purpose registries, including the state public health surveillance and notifiable condition registries; identifying deceased patients by linking mortality data from federal and state sources to the health information exchange global master patient index; and linking public health surveillance data to mortality data to identify potential factors correlated with mortality.

Scope

Background

Scattered data hinders many healthcare processes. Healthcare information is increasingly 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 for which there may be multiple identifiers as well as for data collected about the same patient at different healthcare institutions, different pharmacy systems, different payers, different public health agencies, and so on. This situation hinders the aggregation of information about individuals across such databases as needed for clinical decision support, clinical care, public health reporting, clinical research, and outcomes management. Aggregation is important not only to determine a patient's healthcare status,but also for clinical effectiveness research, drug safety research, and other population-based studies requiring comprehensive data.

Context

Although HIEs are an increasingly common source of comprehensive clinical data, formal recommendations explicitly addressing HIE data aggregation approaches are lacking. Consequently, HIEs currently use a variety of differing data aggregation approaches. Furthermore, although the Office of the National Coordinator for Health Information Technology (ONC) has initiated a national effort to disseminate health interoperability specifications that include patient identity management transactions, the specifications are silent with respect to specifications for data aggregation methodologies and algorithms. Because HIEs 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 data aggregation are crucial to developing a robust National Health Information Network (NHIN).

Settings

The data analyzed is part of an existing data repository created by Regenstrief informaticians. The INPC is a 15+ year-old operational and secure health information exchange. It is a centrally managed, federated, clinical data repository that supports a variety of services, and it is a long-term partnership among the major Indianapolis hospital systems. The health information exchange includes a broad array of participants and members, including competing health systems, physician practices, laboratories, imaging facilities, public health agencies (Indiana State Department of Health and Marion County Health Department), payers, Indiana Office of Medicaid Policy and Planning, and researchers. Currently, the majority of the INPC participants are located in the metropolitan Indianapolis area and its collar counties in Central Indiana; however, participation has recently expanded to the northwestern, north central, and southwestern areas of the state. The INPC data repository carries more than 4.5 billion pieces of clinical data, including 50 million text reports; more than 80 million radiology images; and 750,000 EKG tracings, and 25 million different patient registrations for approximately 12 million unique patients.

Methods

Study Design

Our overall approach was to evaluate both fuzzy match and probabilistic data aggregation methods used in a long-standing, operational health information exchange. This characterization included assessing their performance characteristics in various common HIE-related data aggregation scenarios. The extended methods were evaluated using data from real-world clinical scenarios and using synthetic data. Because the quality of data and the variety of data types available for matching influence data aggregating results, we implemented methods to assess HIE data for use in linkage. Finally, an important measure gauging the value of data aggregation methods is studying their utility in real-world comparative effectiveness and health services research. To that end, we applied these data aggregation methods to real-world scenarios to evaluate a variety of healthcare processes and outcomes.

To evaluate various classes of linkage methodologies currently used in the context of a long-standing, operational health information exchange, including a deterministic fuzzy-match linkage algorithm used for unsupervised, real-time linkage, and a probabilistic linkage algorithm with novel modifications also used for unsupervised linkage, we deduplicated a public health registry and linked two separate hospitals within the information exchange to determine overlap. For these scenarios, we manually reviewed a sample of the data set to establish the truth set for additional evaluation. These data sets supported multiple evaluations for this initiative.

To evaluate extensions to the probabilistic linkage method that were designed to improve algorithm accuracy by overcoming shortcomings in existing probabilistic linkage models, we implemented stochastic and closed-form approaches for parameter estimation methods; generalization of the probabilistic method was used to accommodate statistical dependence between fields.

To assess the data characteristics of specific data fields that inform the selection of specific data elements for matching, we implemented metrics. These metrics include Shannon's entropy, theoretical maximum entropy, percentage of theoretical maximum entropy, unique values, average token frequency, null values, null value rate, collision rate, and number of potential pairs.

To evaluate and characterize technical performance and clinical and operational value of linking real-world HIE data sources, comparing the effectiveness of both deterministic and probabilistic methods, we deduplicated a public health registry, linked hospitals within the HIE, linked patients across more than 90 hospitals in the INPC, and linked mortality data (death status) to INPC data.

Data Sources/Collection

The data analyzed is part of an existing data repository created by Regenstrief informaticians. The INPC is a 15+ year-old operational and secure health information exchange. It is a centrally managed, federated, clinical data repository that supports a variety of services and is a long-term partnership among the major Indianapolis hospital systems. The health information exchange includes a broad array of participants and members, including competing health systems, physician practices, laboratories, imaging facilities, public health agencies (Indiana State Department of Health and Marion County Health Department), payers, Indiana Office of Medicaid Policy and Planning, and researchers. Currently, the majority of the INPC participants are located in the metropolitan Indianapolis area and its collar counties in Central Indiana; however, participation has recently expanded to the northwestern, north central, and southwestern areas of the state. The INPC data repository carries over 4.5 billion pieces of clinical data, including 50 million text reports; more than 80 million radiology images; and 750,000 EKG tracings, and 25 million different patient registrations for approximately 12 million unique patients. We also studied a public health client registry containing more than 750,000 patient records. Additional specific details can be found in the peer-reviewed publications.

Interventions

This research relied exclusively on existing electronic medical records. The patients were not contacted or intervened on in any way.

Measures

A variety of outcome measures were reported in our peer-reviewed publications. These measures included assessments of accuracy, including sensitivity, specificity, positive predictive value, and receiver-operator area under the curve (AUC) measures. Goodness of model fit measures were also reported and included chi-square analysis, AIC, and BIC. Additional details can be found in the peer-reviewed publications.

Limitations

The results of this initiative apply to patient data gathered from heterogeneous clinical data sources. The algorithms evaluated included the ubiquitous Fellegi-Sunter probabilistic method and a custom heuristic algorithm internally developed at the Regenstrief Institute. To the extent that other data sources reflect heterogeneous data and leverage similar probabilistic or heuristic methods, results will be maximally generalizable.

Results

Principal findings

Parameter estimation: We observed that, for cases in which stringent blocking criteria produce a high match prevalence (e.g., > 90%), the EM estimation method cannot distinguish between matches and non-matches in part because of the lack of a representative cohort of non-matching candidate pairs. To accommodate this variation, we developed approaches to estimate the u parameters empirically and then fix them while using the EM to maximize the remaining parameters (m and P). We assessed three approaches to estimating the u values empirically, including: (1) randomly sampling within the block and (2) using all the data from the block to estimate the probability of chance agreement between fields (closed form within blocks) or (3) to randomly sample from the entire population to estimate chance agreement. Our results indicated that estimation of u parameters by using a random sample from within the block is the most efficient computationally and provides u estimates most reflective of the blocked data. We conclude that estimating the u parameters using EM is ideal; however, if the approximate match prevalence is unknown and the initial bootstrap estimates for match prevalence are inaccurate, then EM may yield inaccurate estimates. Thus, estimating u parameters using random sampling from within blocks may produce more accurate estimates for the records under consideration.

Accommodating independence: We developed a methodology to identify field dependence that should be accounted for when using a latent class record linkage model. We also describe how this dependence can be incorporated into the Fellegi-Sunter model. The proposed method was shown to produce a statistical model with overall better model fit.

Optimizing record pair review: To improve the efficiency of reviewing record pairs, which establish a reference gold standard, we have developed a novel approach using McNemar's test, allowing fewer records to be analyzed when comparing two linkage models. This method informs synthetic linkage data that are derived from real-world data for which we are developing processes to create prior probabilities and representative comparator distributions for key data characteristics for alphanumeric data types. We developed methods for summarizing each data element and vector-based match-type and are validating components of this model. In addition to the McNemar's test, we developed a novel approach to optimal record pair sampling when comparing different linkage methods. The sampling approach can substantially reduce variance by over sampling pairs with discordant results and under-sampling subjects with concordant results. We developed and published a heuristic sampling rule for when there is no prior knowledge of the individual sensitivities and specificities, or the true prevalence of the true positive findings, in the study population.

Improving model fit: Commonly used linkage models assume that agreement patterns among multiple fields within a latent class are independent. This "conditional independence" assumption often is not valid. We developed and evaluated a novel methodology to identify and accommodate field dependence when using a latent class model. This method produces a better model fit.

Manual review process: We refined our processes for manually adjudicating record pairs to establish an analytical reference data set. A recent finding is that 1) we've discovered that different human reviewers demonstrate substantial variation when determining true match status; 2) because of this consistently observed variation among multiple reviewers, we conclude that each record pair in a reference data set must be adjudicated by multiple reviewers (at least two) to identify and resolve subjective reviewer discrepancies; 3) much of the patient matching literature is based upon single-adjudicator data sets, 4) so this finding calls for greater scrutiny when assessing the accuracy of the body of extant record linkage literature.

Probabilistic modeling: The F-S model makes strong assumptions that are possibly violated in real-world applications. However, relaxing these assumptions with more flexible models may not always improve the classification accuracy, although the statistical model fit improves substantially. In general, if the assumptions for the F-S model are violated, it is worthwhile to pursue the more complex models, such as log linear model and Gaussian random effects model, for improved classification accuracy only when the match prevalence is close to 0 or 1. Match classification accuracy is influenced by 1) extreme match prevalence (near 0% or 100%), 2) the degree to which the independence assumptions are violated (strong or weak conditional dependence), and 3) how much discriminating power individual fields have. It is

important to select fields with high discriminating power and discard fields that do not significantly contribute to the classification of the record pairs. This is because fields with high discriminating power can decrease the bias that the F-S model provides when match prevalence is close to 0 or 1 and conditional dependence is present.

Outcomes

Assessing field characteristics: We developed algorithms (using the open-source R statistical system) that create metrics for each field and pairwise field combinations. These metrics characterize discriminating power and information content for each matching field in a data set. Metrics we have implemented include information entropy (H), maximum entropy, null rates, and random collision. We established baseline metrics across several data sources in the Indiana Network for Patient Care. These baseline measures will inform a feature selection method for optimizing blocking strategies for matching. These metrics have been incorporated into a patient deduplication module found at https://modules.openmrs.org/modules/view.jsp?module=patientmatching, which is part of the open source electronic medical record system OpenMRS (www.openmrs.org).

Analysis of the Regenstrief matching system used in the INPC was included in invited testimony presented before the Office of the Coordinator for Health Information Technology's Privacy and Security Tiger Team hearing, held on December 9, 2010, in Washington, DC.

Evaluating linkage using public health data: We have worked with our partners at the Marion County Health Department to link and deduplicate their master patient registry, which contains data from a variety of public health initiatives, including Women, Infants, and Children (WIC); immunization programs; lead screening programs; and notifiable condition reporting initiatives. To integrate data from such a variety of processes, we developed a data quality and evaluation process that identifies and corrects common formatting issues. We worked with our partners at the Indiana State Department of Health to link syndromic surveillance cases from across the state in the manner to identify patients who frequently use emergency department services, assess crossover patterns of patients among emergency departments, and evaluate factors that predict outcomes in the emergency department. A manuscript describing the results of these analyses received the America Medical Informatics Association 2011 Fall Symposium Distinguished Paper award and received an award for "Outstanding Research Article in Biosurveillance" from the International Society for Disease Surveillance in August 2012. Additionally, we showed that patient matching techniques developed in this project were able to successfully support evaluation the impact of selective mapping strategies on automated laboratory result notification. Matching methods developed under this project were successfully evaluated for their ability to support integration and analysis of public health reportable disease data. That analysis was published in the AMIA 2012 Fall Symposium and received the American Medical Informatics Association 2012 Fall Symposium award, “Best of AMIA in Public Health Informatics,” for “An Evaluation of the Rates of Repeat Notifiable Disease Reporting and Patient Crossover Using a Health Information Exchange-based Automated Electronic Laboratory Reporting System.” Matching methods developed under this project supported the development of a predictive model that accurately identifies patients who frequently use emergency department services, to be presented at the AMIA 2013 Fall Symposium, and a manuscript that is currently under peer review.

Open Source: Dr. Grannis collaborates with a freely available, open-source, community-developed patient matching software project called “OpenEMPI” (www.openempi.org). Best practices learned from the analyses in this initiative have informed the design of OpenEMPI. In September 2012, under the leadership of Dr. Grannis, OpenEMPI was deployed as the healthcare master person index for the country of Rwanda.

Discussions

Please see peer reviewed manuscripts for more discussion.

Conclusions

Current record-matching approaches can be improved. Though we have demonstrated several optimization and improvement to both patient matching algorithms and process for managing data used for record linkage, additional

improvements and best practice recommendations are needed to achieve the goal of an effective learning healthcare system.

Significance

Patient identification and record linkage are key components of the emerging learning healthcare systems, of which thr lifeblood is integrated electronic data. Recommendations for record linage best practices are needed. Our findings contribute to the body of record linage knowledge and reprint approaches to optimize record linkage outcomes.

Implications

Through a variety of peer-reviewed record linkage analyses, we have demonstrated that probabilistic record linkage is accurate and can be feasibly applied to a variety of data sets, ranging from clinical to population based. There is a paucity of evidence-based best practices for patient matching approaches. Our work provides evidence-based guidance for accurately and efficiently sampling record pairs to establish "gold-standard" reference data sets upon which accuracy measurements can be made.

List of Publications and Products

Daggy, J. K., Xu, H., Hui, S. L., Gamache, R. E., & Grannis, S. J. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. BMC Medical Informatics and Decision Making. doi: 10.1186/1472-6947-13-97

Daggy, J. K., Xu, H., Hui, S., & Grannis, S. (currently in revision). Latent Class Models with Conditional Dependence in Record Linkage. Statistics in Medicine.

Daggy, J., Hui, S., Gamache, R.E., Grannis, S.J. (under revision). Studying What Works in the Real- World: Identifying Best Practices for Parameter Estimation of Key Elements in Probabilistic Record Linkage. Submitted manuscript to AMIA 2011 Annual Symposium, but was not accepted.

Finnell, J.T., Overhage, J.M., Grannis, S. (2011). All Health Care is Not Local: An Evaluation of the Distribution of Emergency Department Care Delivered in Indiana. AMIA 2011 Annual Symposium Proceedings.

Gamache, R.E., Dixon, B.E., Grannis, S., Vreeman, D.J. (2012). Impact of Selective Mapping Strategies on Automated Laboratory Result Notification to Public Health Authorities. AMIA 2012 Annual Symposium Proceedings.

Gichoya, J., Gamache, R.E., Vreeman, D.J., Dixon, B.E., Finnell, J.T., Grannis, S. (2012). An evaluation of the rates of repeat notifiable disease reporting and patient crossover using a health information exchange-based automated electronic laboratory reporting system. AMIA 2012 Annual Symposium Proceedings. PMCID: PMC 3540527.

Grannis, S. (2010). Grannis Patient Linking Testimony 2010_12_09. Written testimony. 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.

Grannis, S., Dixon, B., Xia, Y., Wu, J. (2012). Using Information Entropy to Monitor Chief Complaint Characteristics and Quality. International Society for Disease Surveillance 2012 Abstract. Control ID: 1516638.

Keiper, J., Kamatham, C., Kankanala, A. (not yet submitted for peer review). Supervised Learning of Record Linkage Field Metrics.

Wu, J., Xu, H., Finnell, J.T., Grannis, S. (2013). A Practical Method for Predicting Frequent Use of Emergency Department Care Using Routinely Available Electronic Registration Data. AMIA 2013 Annual Symposium Proceedings.

Xu, H., Hui, S. L. and Grannis, S. (2013), Optimal two-phase sampling design for comparing accuracies of two binary classification rules. Statist. Med.. doi: 10.1002/sim.5946