Enhancing Patient Matching in Support of Operational Health Information Exchange

Principal Investigator: Shaun J. Grannis, MD, MS

Team Members: Huiping Xu, PhD, Xiaochun Li, PhD, Suranga Kasthurirathne, PhD, Josh Vest, PhD, MPH, Chris Harle, PhD, Bo Na, MS, Andrew Martin, Lauren Lembcke, Kelly Mosesso, Ashley Griffith, MHA, Jennifer Williams, Amanda Woods

Organizations: Trustees of Indiana University, Regenstrief Institute, University of Florida

Inclusive Dates of Project: July 1, 2017 – April 30,2023

Federal Project Officer: Jesse Crosson

This project was funded under grant number 5R01HS023808-05 from the Agency for Healthcare Research and Quality (AHRQ), U.S. Department of Health and Human Services (HHS). *The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of AHRQ. Readers should not interpret any statement in this report as an official position of AHRQ or of HHS. None of the authors has any affiliation or financial involvement that conflicts with the material presented in this report.*

Structured Abstract

Purpose: This study sought to enhance the accuracy of linkage methods for integrating patient data across data sources. The primary objectives were to compare methods, evaluate methodological enhancements, and assess the impact of data standardization and normalization on match accuracy.

Scope: The study used four gold standard matching datasets containing approximately 62,000 records from patient matching use cases. Patient records from the Indiana Network for Patient Care (INPC), with over 47 million patient records from 100 clinical sources, were used to evaluate methods. The study encompassed various use cases, including deduplication of HIE patient registry records, linking public health client registries, newborn data deduplication, and linking vital records to ascertain death status.

Methods: The research applied linkage methods to clinical data and assessed the impact of several enhancements, including handling missing data, considering conditional dependence, incorporating nearness comparison, and applying data standardization. Manual review of record pairs was conducted to establish a gold-standard match status.

Results: Key findings included the value of token frequency in matching, the importance of accounting for conditional dependence, and the benefits of data standardization and similarity measures. Handling missing data using the missing at random method significantly improved match accuracy, particularly for sensitivity and F score.

Conclusion: This research sought to improve linkage accuracy to integrate fragmented patient data, employing probabilistic approaches, considering missing data, and utilizing various enhancements. These findings highlight the potential for more accurate patient data integration in healthcare settings.

Keywords: probabilistic and deterministic linkage, conditional dependence, data standardization, missingness, match accuracy.

Purpose

The objective of this study was to implement emerging recommendations for matching data enhancements in combination with novel matching algorithms enhancements and measure the resulting matching accuracy improvements. Such evidence-based outcomes can inform future formulations of the national patient identity management strategy. We accomplished this goal with the following specific aims:

Specific Aim 1: We implemented three general classes of recommended matching data enhancements and measured the resulting matching accuracy improvements. Using the four gold standard patient matching datasets, each of the 3 recommendations were evaluated independently and in combination by comparing enhanced matching data results to baseline matching results derived from the original unmodified datasets.

Specific Aim 2: We implemented four novel matching algorithm enhancements and assess the resulting matching accuracy improvements. Using the four gold standard patient matching datasets, we assessed the effectiveness of each algorithm modification independently and in combination by comparing enhanced matching algorithm results to baseline matching results derived from the original unmodified Fellegi-Sunter algorithm.

Specific Aim 3: We measured the matching accuracy improvements resulting from using combinations of (a) three best practice matching policy recommendations and (b) four novel matching algorithm enhancements. Using the four gold standard patient matching datasets, we assessed the effectiveness of each remaining combination not evaluated in aims 1 and 2 by comparing enhanced matching results to baseline matching results derived from the original unmodified dataset and unmodified algorithm.

Scope

Significance

1. Health care data is fragmented. Patient information is fragmented across the US health care system. Each encounter with a hospital, health system, outpatient provider, clinic, pharmacy, long term care provider, or public health agency, generates information. Problematically, this information is stored in numerous independent clinical repositories with no single unique identifier to enable an integrated comprehensive patient record[1,2]. Even within single large institutions, like a health system or hospital, the internal billing systems, laboratory information systems, and electronic health records are independent silos of information using different identifiers, which may or may not exist in other systems. This fragmented information risks patient safety, hinders data aggregation for clinical decision support, prevents physicians from having comprehensive medical information, deters effective population health approaches, creates inefficiencies by delaying care, limits public health reporting, and severely reduces the utility of electronic patient information for clinical research. Accurate patient matching, which is defined as identifying records for the same person across separate systems, is necessary for the delivery of safe and effective health care and to realize the nations' cost and quality improvement goals[3,4].

2. Improved algorithms and enhanced matching data can decrease data fragmentation and improve matching accuracy. Optimizing patient matching algorithms and enhancing data used for matching are our best approach to integrating patient records across disparate systems. Matching approaches such as a national patient identifier (NPI) or biometric identifiers are infeasible due to political, financial, and operational barriers. While *a*n NPI can improve matching efficiency and accuracy[5,6], implementation is estimated to cost between $1.5 and $11 billion over several years. More importantly, ongoing privacy and security concerns related to the widespread sharing of patient information have hindered the implementation of an NPI[7]. Congress suspended federal funding for an NPI due to privacy concerns in 1999 and although states and other entities are not prohibited from implementation, there is no political interest in NPIs in the US. Likewise, biometric approaches such as fingerprints, iris patterns, facial shapes and vein patterns can significantly improve matching[8]. However, biometrics are expensive and biometric identifiers can evolve[9]. Voice patterns can change gradually with age or abruptly with illness; fingerprints can degrade (disappear) with time; retinal patterns can change in patients with conditions that affect eye, such as diabetes. Additionally, privacy concerns remain over the co-opting of biometric identifiers for additional non-health care uses such as fingerprint use by law enforcement agencies[10]. Biometrics are not currently widely deployed in the US health care system primarily due to cost and perceived privacy concerns. While these approaches may improve the matching process, they are unlikely to be true solutions in the near-term and are still not panaceas. Robust patient matching methods would still needed be needed to account for uneven adoption across the health care system, to link historical data, to link episodes of care when NPI cards were lost or biometric scanners were down, and to identify inevitable duplicate patients.

3. Evidence-based approaches for optimizing matching algorithm accuracy are needed. Patient matching approaches include a spectrum of increasingly sophisticated methodologies: deterministic, fuzzy match, and statistical/machine learning methods. The simplest approach for matching records compares selected data elements such as name, birth date, gender and Social Security number using exact or deterministic rules[11,12]. Although these algorithms are generally simple to implement and achieve excellent specificity, they are inflexible to changing data characteristics and can lack discriminating power. Intermediate algorithms employ fuzzy logic to address nicknames, typographical errors[13,14], and implement ad-hoc scoring systems using simple heuristics. These represent a reasonable middle ground for performance and implementation complexity, but lack the flexibility of more advanced algorithms. The most sophisticated patient matching approaches leverage statistical and machine learning models to establish match status, and these more complex methods generally yield greatest accuracy. Example models include Bayesian algorithms[15], maximum entropy algorithms[16], and the Felligi-Sunter (FS) maximum likelihood algorithm[17]. With its ability to improve accuracy as more data is presented, and adapt as the data characteristics evolve, the FS algorithm is a core component of many probabilistic matching algorithms in use today.

4. Evidence-based approaches for enhancing data used for matching are needed. While algorithms can be quite effective when implemented properly, algorithms must be paired with high quality discriminating data to maximize matching accuracy. The evidence base guiding optimal implementation for combinations of algorithm is lacking[18]. There have been few formal, comprehensive evaluations of consensus-based matching strategy recommendations using real-world, heterogeneous health care data. Studies often use data derived from small numbers of sources, sources with similar data characteristics, or similar workflows. Further, some evaluations are biased by including human supervision to train and maintain algorithm accuracy. The amount of human supervision required to achieve the level of accuracy described in published studies may be infeasible for most health care data repositories that capture hundreds of thousands of new clinical records daily.

5. Expert panel recommendations for matching are emerging. The recognized lack of consistent matching practices has generated multiple best practice recommendations, but no consensus. Several organizations including the Agency for Healthcare Research and Quality (AHRQ)[19], the Health Information Management Systems Society (HIMSS)[20], the Bipartisan Policy Center[21], the eHealth Initiative[22], the Markle Foundation[23], and the Office of the National Coordinator for Health IT (ONC)[24,25] have either promoted or published best practice recommendations for patient matching. However, there is a paucity of peer-reviewed research specifically addressing the feasibility, effectiveness, and generalizability of implementing specific recommendations. Supporting this notion, a March 2014 U.S. Government Accountability Office (GAO) report evaluating HIE and patient matching notes, "HHS developed an electronic health information exchange strategy that includes principles to address key challenges but lacks specific prioritized actions."[26]

6. Lack of evidence for the effectiveness of expert recommendations can hinder adoption. With an incomplete evidence base unable to more firmly support existing guidance, stakeholders may be less inclined to pursue approaches with unclear value, or they may implement methods that upon further study prove to be less effective and generalizable than initially perceived. Thus, our proposal motivated by the dubious state of the strategy for patient identity management in the US. We recognize that approaches for consistent, accurate and efficient patient identification are needed for a fully functional learning health care system, yet questions remain as to whether we can achieve acceptable, consistent matching performance[5,6].

7. We evaluated best-practice matching data recommendations and novel algorithm improvements in the context of the nation's most comprehensive health information exchange. The Indiana Network for Patient Care (INPC) carries over 5 billion pieces of clinical data, for approximately 27.5 million unique patient registrations, covering approximately 13.4 million unique patients across more than 100 health care institutions. Using this laboratory we compared baseline matching accuracy to matching results that implement various combinations of a) best practice recommendations for matching data, and b) algorithm enhancements.

Innovation

This work is innovative because it is designed to evaluate the independent and combined effects of two distinct but related strategies, both necessary for advancing the state-of-the-art in patient matching: a) best practice recommendations for data curation policy and process improvements, and b) matching algorithm enhancements.

1. Matching data enhancements. Regarding best practice recommendations for data curation policy and process improvements, we implemented and evaluate data preprocessing enhancements, including: matching data standardization, augmenting data to increase discriminating power, and evaluating and improving data quality. These best practice recommendations have not been evaluated in the context of a large, heterogeneous health information exchange.

2. Matching algorithm enhancements. In addition to policy and process enhancements, our experienced data analytics team also implemented and evaluate novel enhancements to commonly used patient matching methods. We propose to extend the current research evidence base by applying four novel, generalizable approaches to improve the accuracy of patient matching. The novel approaches draw on probabilistic and machine learning methods and include: extending patient matching models to leverage value-specific frequencies for key matching fields, incorporating similarity metrics into agreement comparators, accounting for agreement correlation among fields, and accommodating missing data.

3. Matching scenarios highlighting different health contexts. The same matching approach may not be optimal for all situations because the data available in different matching scenarios exhibit varying degrees of discriminating power and data quality. In this study, we evaluated patient matching strategies in the context of four use cases with significant clinical, public health, and research implications. For each use case we used manually reviewed gold standard patient matching data sets derived from the HIE. All data sets have supported prior peer reviewed patient matching research. A brief description of each data set follows, and is further described in Table 1.

4. Newborn screening[27,28]. Not all infants are appropriately screened for harmful or potentially fatal disorders that are otherwise unapparent at birth[29]. 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[30]. To improve identification of unscreened infants, we developed an algorithm to link records from Indiana's statewide newborn screening registry to the INPC[31]. For this analysis we extracted newborn screening lab records and INPC records for patients less than 1 month of age, randomly sampled and manually reviewed over 11,000 record pairs to create a gold standard analytic data set.

5. Linking hospital patient registries[32,33]. We examined our matching enhancements in the context of linking enterprise master patient registries from hospital systems that share overlapping patient populations. As hospital systems are increasingly incentivized to partner through ACOs and other value-based purchasing models[34,35,36], linking their records for common patients was a critical first step in more coordinated and lower cost care. To routinely evaluate the HIE matching accuracy we have created a randomly sampled, manually reviewed data set containing over 16,000 record pairs for health systems sharing the same patient.

6. Removing duplicates public health client registry[18,37]. Public health registries help track the health trends of populations and support many public health activities. Data in these registries derive from multiple public health service areas and exhibit varying data quality. We de-duplicated the complete patient registry for the Marion County Health Department (MCHD), Indiana's largest public health department. De-duplication is a class of record linkage where a data set is linked to itself to identify potential duplicate records. To evaluate the accuracy of the de-duplication process, we manually reviewed over 17,000 record pairs to create a gold standard evaluation set.

7. Ascertaining death status[38,39]. We examined matching enhancements in the context of social security death data linked to many participating organizations in the INPC. Accurately and comprehensively updating health records with patients' accurate vital status is critical to robust clinical quality measurement, public health reporting requirements, and high quality clinical research. To evaluate these matching approaches, we linked INPC hospital registry records to the Social Security Death Master file (SSDMF) and manually reviewed over 12,000 randomly sampled record pairs to create a validated data set for analysis.

Overall, this project was innovative because we used unparalleled health information exchange data to evaluate the effectiveness of recommended but poorly understood matching strategies. We conducted this evaluation in the context of four clinical and public health use cases that are of current and growing importance to improving health care quality and outcomes while reducing costs.

Table 1: List of gold standard, peer reviewed record linkage datasets for us in evaluation. The variety of clinical and population health scenarios represent important clinical use cases, and also reflect a spectrum of data characteristics that pose challenges to accurate linkage, including varying data quality and discriminating power.

Dataset	Linkage Description	Data Characteristics	Use Case Value
Newborn	Statewide public health newborn screening laboratory results liked to HIE clinical records	Missing names (infants), richer parent demographics, few unique identifiers, varying completeness, and standardization	Public health and pediatric clinical stakeholders seek to improve neonatal care management
Health System	Patient registries from different health systems that share overlapping patient populations	Varying completeness and standardization, unique identifiers present.	Health systems seek to improve care coordination, population health management
Public Health	Public health population registry covering a large metropolitan area	Inaccurate and incomplete patient data, few unique identifiers.	Public health seeks to improve data quality for more effective public health management
Data Registry	National social security death data linked to all participating HIE organizations	Unique identifiers (SSN) present, high completeness.	Public and health systems seek to improve quality of mortality data to assess care outcomes; improve tumor registry data

8. AHRQ Priority Populations. With nearly total state population coverage, the INPC catchment population includes women, children, racial and ethnic minorities, populations with special health care needs (chronic illness, disabilities, and end of life care needs), the elderly, low-income, inner-city, and rural populations.

Methods

Using randomly sampled, manually reviewed gold standard datasets derived from four distinct clinical patient matching use cases, we evaluated matching accuracy improvements resulting from implementing a) best practice recommendations for data curation policy and process improvements, and b) matching algorithm enhancements through three specific aims. For each aim we used the traditional FS patient matching algorithm to measure the relative improvements in matching sensitivity, specificity, positive predictive value (PPV), and the area under the ROC curve (AUC) compared to baseline matching approaches without incorporation of the specific matching enhancement. Each matching enhancement was applied to all four gold standard matching datasets. Table 2 highlights the project's overall approach.

In aim 1 we studied matching accuracy improvements resulting from enhanced matching data generated using three best-practice policy recommendations for data curation and standardization. Using the four gold standard patient matching datasets, each of the 3 recommendations were evaluated independently and in combination by comparing enhanced matching data results to baseline matching results derived from the four original unmodified gold-standard datasets.

In aim 2 we evaluated matching accuracy improvements resulting from four novel matching algorithm innovations. Using the four gold standard patient matching datasets, we assessed the effectiveness of each algorithm modification independently and in combination by comparing enhanced matching algorithm results to baseline matching results derived from the original unmodified FS algorithm.

In aim 3 we evaluated matching accuracy improvements resulting from combining both best practice recommendations and algorithm enhancements. Using the four gold standard patient matching datasets, we assessed the effectiveness of each remaining combination not evaluated in aims 1 and 2 by comparing enhanced matching results to baseline matching results derived from the original unmodified dataset and unmodified algorithm. We excluded any best practice recommendations and algorithm enhancements that failed to showed significant improvement in aims 1 and 2.

Table 2: Breakdown of analyses by matching enhancement type and specific aim. 'D' refers to combinations of the three "Data" enhancements. 'A' refers to combinations of the four "Algorithm" enhancements. Each cell represents a combination of 'Data' and 'Algorithm' enhancements to be analyzed. Each analysis (represented by a cell in the table) will measure sensitivity, specificity, positive predictive value, and AUC using 4 validated matching datasets. Each cell represents 16 analyses (4 accuracy measures x 4 datasets). The maximum number of analyses will be: ((8 rows x 16 columns)- 1) x 16 analyses/cell = **2,032 analyses.** Aim 1 will have (7 cells x 16 analyses/cell) = 112 analyses. Aim 2 will have (15 cells x 16 analyses/cell) = 240 analyses, and Aim 3 will have up to (105 cells x 16 analyses/cell) = 1,680 analyses.

ALGORITHM ENHANCEMENTS

DATA ENHANCEMENTS

	NONE	A1	A2	A3	A4	A1,A2	A1,A3	A1,A4	A2,A3	A2,A4	A3,A4	A1,A2,A3	A1,A2,A4	A1,A3,A4	A2,A3,A4	A1,A2,A3,A4
NONE		AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2	AIM 2
D1	AIM 1	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3
D2	AIM 1	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3
D1,D2	AIM 1	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3
D1,D3	AIM 1	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3
D2,D3	AIM 1	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3
D2,D3	AIM 1	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3
D1,D2,D3	AIM 1	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3	AIM 3

Specific Aim 1: Implement three general classes of recommended matching data enhancements and measure the resulting matching accuracy improvements.

Incomplete and incorrectly attributed patient information can lead to suboptimal and inappropriate care[40,41] Best-practice recommendations exist to improve patient matching algorithms and approaches. However, despite the development of these recommendations, best-practice suggestions for improving patient matching performance are still very much in the emergent stage, and their collective (and individual) effectiveness has yet to be determined. This aim assessed which current set of recommendations creates the greatest improvements in patient-matching algorithm performance.

We *hypothesized* that implementing the proposed best practices for matching, including (a) adding more matching fields, (b) standardizing data, and (c) improving data quality will improve matching sensitivity, specificity, positive predictive value, and AUC. To test this hypothesis, we evaluated each recommendation separately and in combination using previously validated gold-standard matched data sets derived from our rich HIE. This evidence meaningfully informed next steps in developing a national patient matching strategy.

Justification. Influential policy and industry organizations representing the interests of health information technology and patient safety have developed policy recommendations for improving patient matching processes. Agencies including AHRQ[19], HIMSS[20], the Bipartisan Policy Center[21], the eHealth Initiative[22], the Markle Foundation[23], and ONC[24,25] have either promoted or published best practice recommendations for patient matching. These recommendations vary and none have been formally evaluated. The recommendations for improving patient matching include three distinct approachess[24]: (a) adding more matching fields to increase discriminating power; (b) adopting uniform field-specific data standardization methods to ensure consistency; and (c) assessing and improving matching field accuracy and completeness to ensure adequate data quality. We describe our approach to implementing these recommendations below.

(a) Additional matching fields can improve discriminating power**.** Additional independent data elements that are routinely available (but not commonly used) can increase discriminating power of matching algorithms. We added candidate matching fields to our validated matching data sets. Additional fields include (1) middle name, (2) mother's and father's first and last names, (3) mother's maiden name, and (4) email address to our existing matching data sets. While these data elements are not routinely used for matching, they are present within transactions shared among our HIE stakeholders[42]. With over 100 hospitals transmitting over 1 million clinical transactions daily to the INPC, we are well positioned to study these candidate fields across a diverse set of health care institutions. For example, a previously published analysis revealed that parents' first and last name were present but not used for matching for 58% of records in a random sample of newborns[27], which suggests these fields have the potential to improve matching accuracy.

(b) Data format standardization methods to ensure consistency. While different systems may use similar matching fields, they're often stored in inconsistent formats. Social Security Number, address, telephone, and even names may be formatted differently (e.g., "(317) 555-1212" versus "3175551212", "O'Brien" versus "Obrien"), resulting in less accurate matching. The ONC recommendations for improving data consistency and normalization point to field-specific standardization processes, including the Council for Affordable Quality Healthcare's (CAQH) name standards rule[43], X12 transaction set standards for demographics[44], International Organization for Standardization (ISO) formats[45], and United States Postal Service (USPS) address standards[46]. We also applied the International Telecommunication Union (ITU) E.123 recommendations for telephone and email address standardization[47]. We applied these standardization rules to our existing matching data sets as described in Table 3).

Table 3: Specific data format standardization recommendations mapped to existing and additional fields.

Additional Fields	**CAQH**	**X12**	**ISO**	**USPS**	**ITU-T**
Mother's Last Name	X	X			
Mother's First Name	X	X			
Father's Last Name	X	X			
Father's First Name	X	X			
Mother's Maiden Name	X	X			
Email Address					X
Existing Fields					
Last Name	X	X			
First Name	X	X			
Middle Name	X	X			
Gender			X		
Date of Birth			X		
Address				X	
City				X	
State				X	
Zip Code				X	
Telephone					X

(c) Improved data quality can improve matching accuracy. Accurate patient matching requires not only robust algorithms but also high quality data[48]. Missing fields, automatically entered "default" values, typographical errors, misspellings, and false information hinder data quality and are associated with inaccurate matching. Strategies to improve data quality often focus on data entry process improvement, data validation, and data cleaning techniques[49]. Because our patient matching laboratory includes participants from over 100 heterogeneous organizations, it is impractical to deploy and study widespread patient registration and data entry process improvement. Therefore, for this particular recommendation we focused our efforts on evaluating the effectiveness of implementing (1) HIE-based data validation and (2) data cleaning methods, where our team has prior success developing and deploying a health care data quality analysis framework[50], as well as implementing technical processes to improve health data quality[51,52]. Validating data quality and cleaning data can improve matching accuracy by identifying specific data shortcomings to be addressed for a given data source[48]. Our approaches for data validation and data cleaning are described below.

Data validation rules were identified by analyzing the idiosyncratic characteristics of patient matching fields. Default values are defined (e.g., name "John Doe"; date of birth "1/1/1901") and such instances are invalidated for use in patient matching to avoid false positive matches. Additionally, certain person traits can be validated using simple rules. Examples of such rules include: month of birth is limited to one of twelve distinct values; day of birth is limited to one of thirty-one values; names may contain the letters A-Z, hyphens, and single quotes, but no other punctuation and no numbers. SSN's and telephone numbers should contain no more than six nines or six zeros in a row. All traits not complying with specific rules are marked as invalid and are not used for patient matching. Invalid address components, including street number and name, city, state and ZIP code can be identified using more sophisticated rules implemented in various software packages[53].

Data cleaning methods can correct minor data errors and impute missing values. Slightly misspelled names can be mapped to valid closely matching names using string nearness measures (e.g., "Snith" → "Smith")[33] and nicknames mapped to canonical names (e.g., "Robbie" → "Robert")[REF]. Both the misspelled and corrected names may be used for matching. Address errors such as misspellings, typographical errors, and nonstandard abbreviations can be similarly corrected using address cleaning software[53]. Further, a significant proportion of missing gender values can be imputed using first names that are closely aligned with one gender (e.g., "Mary" implies "female"; "John" implies "male"). Missing values can also be derived from historical values maintained within the HIE. When data (e.g., mother's maiden name) is missing for a current demographic record (e.g., the patient may not have provided it during the most recent visit), a prior historical record with that value may be retrieved. We developed and implemented software to provide data quality validation and data cleaning methods using the framework described in Table 4.

Table 4: Approach to applying data validation and data cleaning methods to existing and new matching fields.

Additional Fields	Data Validation	Data Cleaning: Connect Minor Errors	Data Cleansing: Impute Missing Data
Mother's Last Name	**Remove Default Values**	**Close Match**	**Historical**
Mother's First Name	**Remove Default Values**	**Close Match/Nickname Standardize**	**Historical**
Father's Last Name	**Remove Default Values**	**Close Match**	**Historical**
Father's First Name	**Remove Default Values**	**Close Match/Nickname Standardize**	**Historical**
Mother's Maiden Name	**Remove Default Values**	**Close Match**	**Historical**
Email Address	**Remove Default Values/Apply Rules**		**Historical**
Existing Fields			
Last Name	**Remove Default Values**	**Close Match**	
First Name	**Remove Default Values**	**Close Match/ Nickname Standardize**	
Middle Name	**Remove Default Values**	**Close Match**	**Historical**
Gender	**Apply Rules**		**Impute from First Name**
Date of Birth	**Remove Default Values/Apply Rules**		
Address	**Remove Default Values/Apply Rules**	**Address Correction Software**	**Historical**
City	**Remove Default Values/Apply Rules**	**Address Correction Software**	**Historical**
State	**Remove Default Values/Apply Rules**	**Address Correction Software**	**Historical**
Zip Code	**Remove Default Values/Apply Rules**	**Address Correction Software**	**Historical**
Telephone	**Remove Default Values/Apply Rules**		**Historical**

Research Design. The relative improvements of each of the above best-practice policy recommendations assessed individually and in combination using each of four validated test data sets. We assessed matching accuracy (including sensitivity, specificity, PPV, and AUC) before and after implementing standardization and quality improvement processes for both existing and new fields. We implemented each of the three recommended data enhancements, applied the Felligi-Sunter probabilistic algorithm to the enhanced data, and compared the matching accuracy of each enhancement individually and collectively to baseline matching performance. Table 2 highlights the specific analyses for Aim 1, where 'D1' represents adding matching fields, 'D2' represents data standardization methods, and 'D3' represents data validation and cleaning methods. Each of the seven combinations were applied to the four gold standard matching datasets, for a total of 28 enhanced matching analysis datasets. We performed four accuracy analyses (sensitivity, specificity, PPV and AUC) per dataset for a total of 112 analyses.

Analysis. We measured the effect of best-practice recommendations (both individually and combined) on patient matching using sensitivity, specificity, PPV, and AUC for each dataset. Sensitivity, specificity, and PPV were evaluated using proportions and 95% confidence intervals. To account for the clustering effect of multiple methods applied to the same record pair and assess the effects of individual and combined data enhancements on matching performance, we performed a marginal logistic regression using generalized estimating equations[54,55]. For each metric we included main effects, two-way interactions, and three-way interactions of the three enhancements in the model to allow differential effects of a factor as other factors vary. The standard errors of the accuracy measures were calculated using the robust sandwich variance estimation methods. The point estimate and the 95% confidence interval of the AUC, as well as the comparison of multiple AUCs, were performed using a nonparametric approach[56]. Comparison of these accuracy metrics between the four unmodified baseline datasets and the 28 analysis data sets corresponding to the seven possible combinations of the three enhancements and four gold standard baseline data sets were performed using a multiple comparison approach with a Bonferroni adjustment.

Sample Size & Power. Sample sizes for the four gold standard record matching datasets ranged from 11,000 to 17,000. These sample sizes provided at least 80% power to detect a minimum of 2% difference in the AUCs of two correlated ROC curves, assuming one AUC was 80% and the other was at least 82% and the prevalence of true matches lied within the range of 5% to 95%, commonly seen in record linkage applications. These two ROC curves were derived based on the two matching algorithms using combinations of the three enhancements. We assumed a moderate correlation of 50% between matching status determined by the two matching algorithms, recognizing that the power increases as the matching algorithms produces more correlated results.

Specific Aim 2: Implement four novel matching algorithm enhancements and measure the resulting matching accuracy improvements.

Current matching approaches often fail to take advantage of the full discriminating power present in matching data. Aggregate and inferred information (or metadata) are routinely available, but largely unused in matching algorithms. The objective of this aim was to determine the impact of various categories of metadata on patient matching algorithm accuracy. We *hypothesized* that incorporating metadata into matching algorithms would measurably improve matching accuracy with well-defined metrics stated below.

Justification. A primary strategy for improving matching algorithm performance is to add more data elements on which to match records[57]. Intuitively finding matches for "John Public" is enhanced by adding data (e.g. "John *Quincy* Public"). While adding more matching data fields seems logical, it has several practical limitations. Problematically, additional data elements are not always readily available. New data elements are often costly to collect due to required personnel training and system modifications, and are not easily incorporated into existing data collection workflows.

Approach. While capturing new data may be challenging, *information about existing data* are routinely available. The term *metadata* refers to aggregate or inferred information about existing data, and can include 1) field frequencies, 2) similarity measures, 3) dependency relationships, and 4) missing data. Although metadata have the advantage of requiring no additional collection efforts by health care organizations, metadata traditionally have been underutilized for patient matching. Nonetheless, metadata represents a potential source of additional data that can provide additional discriminating power, ultimately improving match accuracy. Below we proposed four approaches for extending current matching algorithms to incorporate several types of metadata to improve patient matching accuracy.

Field-specific String Frequencies. Matching algorithms often use binary inputs (0/1) indicating whether corresponding fields agree. However, these models often do not recognize that agreement may convey varying levels of evidence depending on the specific field value. For example, last name agreement on "Smith" may convey less match certainty than last name agreement on "Harezlak". To capture the varying discriminating power of field-specific values, we extended the base FS model by incorporating an extension to the expectation maximization (EM) algorithm, which is a related statistical method commonly used to parameterize the FS model[58]. Rather than using a binary variable for agreement (1) and non-agreement (0), we used a multinomial representation by stratifying agreements into multiple categories based on individual field strings. Specifically, assume K unique strings for field w (e.g., patient last name) in files of patient records A and B. Combining these with the non-agreement result, there are a total of K+1 possible categories for the given field. Assume that

$$m_{jk} = P\big(a_{ij} = b_{ij} = w_k, a_i \in A, b_i \in B \big| M_i = 1\big), u_{jk} = P\big(a_{ij} = b_{ij} = w_k, a_i \in A, b_i \in B \big| M_i = 0\big),$$

where a_{ij} and b_{ij} are the values in field j for the i$^{\text{th}}$ record pair (a_i, b_i) and $M_i = 0$ or 1 is true match status ($i = 1, 2, \cdots, n, j = 1, 2, \ldots, J, k = 1, 2, \ldots, K$). The match prevalence is denoted as $\rho = P(M_i = 1)$. After combining fields into multiple categories based on distribution of occurrence frequencies, we used the EM algorithm to estimate the parameters and compute the posterior probability, which was used to determine the match status of record pairs.

Similarity Measures. The extension of field-specific string frequencies to the FS model as described above assumes exact field agreement. However, it is not always optimal to compare two strings character-by-character due to typographical error or spelling variations. Rather than treating agreement as binary variable, string comparators measure similarity between two strings by producing a continuous value (ranging from 0 to 1)[33,59,60]. Higher values signify greater similarity, with 1 indicating exact match. A common approach is to dichotomize this continuous similarity measure at a pre-specified cut point. However, incorrectly chosen similarity thresholds can produce inaccurate match results. Identifying optimal similarity thresholds requires assessment of thresholds across multiple fields, and dichotomizing a continuous variable decreases the overall discriminating power[61,62]. We incorporated continuous similarity measures into our base FS matching model using the multinomial approach proposed for incorporating string occurrence frequency. Fields with minimal information content such as gender and middle initial yield binary similarity measures (0/1). Thus, we only applied this method to fields with sufficient information content (as measured by Shannon's entropy to yield a heterogeneous distribution of similarity measures. The EM algorithm developed for the multinomial data for the extended FS model was similarly used to obtain parameter estimates.

Similarity measures and frequency weights can be incorporated into the model both as two separate attributes of the same field, or combined by adjusting the similarity measure by multiplying the minimum of the inverse of frequency weights $g(w_k)$, and scaling the modified similarity interval to (0, 1). Table 6 illustrates the combined adjusted similarity measure calculation. Note that last name "Harezlak" is relatively rare in both files, and so modified similarities of pairs 3 and 4 stay high; however, for the more common name "Smith", the similarities of record pairs 1 and 2 are decreased. The posterior probability of being a match is decreased for record pairs 1 and 2, but increased for record pairs 3 and 4.

Table 6: An example combining both string similarity and token frequency

	Last Names	Similarity	p_A	p_B	Modified Similarity
1	Smith, Smith	1.000	0.9	0.8	0.222
2	Smith, Smth	0.800	0.9	0.001	0.178
3	Harezlak, Harezlak	1.000	0.1	0.2	1.000
4	Harelak, Harezlak	0.875	0.001	0.2	0.875
5	Smith, Harezlak	0.000	0.9	0.2	0.000

Conditional Dependence Among Fields Using Random Effects. The base FS model assumes that field agreements are statistically independent (i.e., not correlated) given a record pair's true match status. For example, for most matching algorithms, knowing that two records agree on (for example) family name does not inform whether other fields (such as birth date) will agree as well. However, in real-world applications this assumption of conditional independence is often violated. For example, agreement status for first name and gender fields are statistically correlated, e.g., two persons with a first name of 'Anna' are likely agree on gender ('Female'). As a result, most matching algorithms fail to leverage such correlations to produce better classification rules[28,63].

To address conditional dependence among fields, we proposed to model record-pair characteristics by introducing an unobserved random variable T into our extended FS models. The random variable T is assumed to follow a standard normal distribution. Given the true match status and the random variable T, the agreement pattern is independent across fields. That is,

$$\begin{aligned} &P\left(a_{ij} = b_{ij} = w_k, a_i \in A, b_i \in B \middle| M_i = 1, T = t\right) \\ &\qquad = \Phi\left(c_{j1k} - \left(a_{j1} + b_{j1}t\right)\right) - \Phi\left(c_{j1,k-1} - \left(a_{j1} + b_{j1}t\right)\right), \end{aligned}$$

for matches and for non-matches, we have

$$\begin{aligned} &P\left(a_{ij} = b_{ij} = w_k, a_i \in A, b_i \in B \middle| M_i = 0, T = t\right) \\ &\qquad = \Phi\left(c_{j0k} - \left(a_{j0} + b_{j0}t\right)\right) - \Phi\left(c_{j0,k-1} - \left(a_{j0} + b_{j0}t\right)\right), \end{aligned}$$

where $w_1, w_2, \dots, w_K$ are ordered from the smallest to the largest based on $\mathrm{g(w_k)}$, $-\infty = c_{jm,-1} \le c_{jm0} = 0 \le c_{jm1} \le \cdots \le c_{jmK} \le c_{jm,k+1} = \infty$ $(m = 0, 1)$ and Φ is the cumulative distribution function of the standard normal distribution. Integrating over the random effect T, we compute the m- and u-probabilities as follows:

$$m_{jk} = \Phi\left(\frac{c_{j1,k+1} - a_{j1}}{\sqrt{1 + b_{j1}^2}}\right) - \Phi\left(\frac{c_{j1k} - a_{j1}}{\sqrt{1 + b_{j1}^2}}\right), u_{jk} = \Phi\left(\frac{c_{j0,k+1} - a_{j0}}{\sqrt{1 + b_{j0}^2}}\right) - \Phi\left(\frac{c_{j0k} - a_{j0}}{\sqrt{1 + b_{j0}^2}}\right)$$

We obtained parameter estimates by maximizing the log-likelihood function via the EM algorithm[64]. When a similarity measure is included we evaluated the random effects model using two approaches. We analyzed the effect of treating similarity measure and frequency weights as (a) two separate attributes of the model and (b) combining the two as a single feature. When large numbers of unique strings are present for a given field, instance data is sparse and the model is heavily parameterized. Thus we parsimoniously combined strings with similar frequency of occurrences to reduce the number of categories and hence the number of parameters. In addition, we evaluated simplifying the random effects approach by focusing on the subset of fields that violate conditional dependence.

Accommodating Missing Data. Data necessary for matching patients is often missing from clinical records for many reasons: values may be unknown, non-existent (a person with no middle name), or omitted due to privacy concerns. For example, pediatricians often don't collect mother's date of birth, even though the information improves record management, because their focus is on the child[65]. Missing fields decrease discriminating power and consequently hinder matching accuracy[66,67,68]. We evaluated three missing data situations.

a. The most restrictive missing data model is the *missing completely at random* (MCAR), which assumes that the missingness of a variable is independent of all observed or unobserved variables. In this situation parameter estimates are unaffected when record pairs with missing data are excluded. However, omitting missing data may lower the precision of estimated parameters due to fewer record pairs being used.

b. *Missing at random* (MAR) is a less restrictive yet more realistic missing data model that assumes the missingness of a variable is independent of the unobserved data, although it can depend on other observed variables. In many situations, MAR represents a reasonable missing data assumption in record linkage.

c. *Missing not at random* (MNAR) asserts that the missingness of a variable is related to the unobserved variable itself. For example, if middle name is absent because it does not exist, a missing value from both records of the record pair can provide information that the two records belong to the same person.

Missing record linkage fields are typically handled by excluding records with missing values on one of the linking fields when estimating match weights[69], or considering the field's agreement pattern as a disagreement[37]. Excluding records with missing values is indicated only when data are MCAR. Thus, excluding records may bias parameters estimates when the MCAR assumption is violated, leading to inaccurate results. Alternatively, treating missing data as disagreement assumes MNAR, which may yield inaccurate results when the MNAR assumption that all missing data represents disagreement is incorrect.

We evaluated the effectiveness of incorporating missing data into our base FS probabilistic linkage model. Representing the FS model using a log-linear approach, we assumed four example fields denoted as A, B, C, and D, and the match status is denoted as M. The FS model can be rewritten as:

$$log(\pi_{ijklm}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_m^M + \lambda_{im}^{AM} + \lambda_{jm}^{BM} + \lambda_{km}^{CM} + \lambda_{lm}^{DM},$$

where π_{ijklm} denotes the probability that a record pair has agreement pattern with i^{th} level of field A, j^{th} level of field B, k^{th} level of field C, l^{th} level of field D, and m^{th} level of match status $(i, j, k, l, m = 0,1)$. Here λ is determined by the constraint $\sum \pi_{ijklm} = 1$ and other $\lambda's$ are parameters satisfying the typical constraints for log-linear models. By using symbols that list the highest order terms for each variable, representation of the FS model is simplified as (AM, BM, CM, DM).

With no missing data, the four binary agreement status variables in fields A through D yield a contingency table with $2^4 = 16$ unique patterns. With missing data, we introduced an indicator for each field having missing values and create an expanded contingency table including both complete and partially observed data. Assuming that fields A and B have missing values, we created two missing indicators, M_A and M_B, and combined them with the four fields to create an expanded table with 2^6 elements. We then incorporated additional terms related to M_A and M_B into the log-linear model to accommodate missing data.

This model used an expanded contingency table enables us to evaluate missing data mechanisms. In the above example, MCAR implies that missing indicators M_A and M_B are independent of all variables and hence the model can be written as (AM, BM, CM, DM, M_A, M_B). When data are MAR, the additional terms for MAR models will only include associations among M_A, M_B and C, D. Further, the MNAR models will include additional terms with associations among M_A, M_B and A, B, or M.

For each unique observed agreement pattern in the expanded table, the probability of this pattern is $\rho\pi_{ijkl1} + (1-\rho)\pi_{ijkl0}$. For patterns not directly observed (e.g., field A is missing), the inferred probability is $\rho\pi_{+jkl1} + (1-\rho)\pi_{+jkl0}$, where $\pi_{+jklm} = \pi_{0jklm} + \pi_{1jklm}$. The log-likelihood for the log-linear model is then $logL = \sum f \cdot log(\pi)$, where f and π are the observed frequency and probability of the agreement patterns, and the sum of the log-likelihood is taken over all unique agreement patterns in the expanded table.

Our approach above described handling the missing data similarly assumes conditional independence of field agreement. This assumption can be relaxed by including interactions among fields in the model. In addition, when similarity measure or frequency weights of a field are available, the proposed multinomial model was used.

Research Design. We evaluated matching accuracy improvements resulting from implementing the four novel matching algorithm innovations using sensitivity, specificity, PPV, and AUC. Using the four gold standard patient matching datasets, we assessed the effectiveness of each algorithm enhancement independently and in combination by comparing enhanced matching algorithm results to baseline matching results derived from the original unmodified FS algorithm. Table 2 highlights the 15 specific algorithm combinations and 240 analyses to be evaluated. Table 2 highlights the specific analyses for Aim 2, where ‘A1’ represents field-specific frequency enhancements, ‘A2’ represents similarity enhancements, ‘A3’ represents methods for conditional dependency enhancements, and ‘A4’ represents incorporating missing data models. Each of the 15 algorithm combinations were applied to the four gold standard matching datasets, for a total of 60 matching analysis datasets. We performed four accuracy analyses (sensitivity, specificity, PPV and AUC) per dataset for a total of 240 analyses.

Analysis. We measured the effect of matching algorithm enhancements (both individually and combined) on patient matching using sensitivity, specificity, PPV, and AUC. Sensitivity, specificity, and PPV was evaluated using proportions and 95% confidence intervals. To account for the clustering effect of multiple methods applied to the same record pair and assess the effects of individual and combined algorithm enhancements on matching performance, we performed a marginal logistic regression using generalized estimating equations[54,55]. For each metric we included main effects, two-way interactions, three-way, and four-way interactions of the four algorithm enhancements in the model to allow differential effects of a factor as other factors vary. The standard errors of the accuracy measures were calculated using the robust sandwich variance estimation methods. The point estimate and the 95% confidence interval of the AUC, as well as the comparison of multiple AUCs, were performed using a nonparametric approach[56]. Comparison of these accuracy metrics between a) the unmodified FS algorithm applied to the four gold standard data sets and b) the 15 algorithm combinations applied to the four algorithm enhancements were performed using a multiple comparison approach with a Bonferroni adjustment.

Sample Size & Power. Sample sizes for the four gold standard record matching datasets ranged from 11,000 to 17,000. These sample sizes provided at least 80% power to detect a minimum of 2% difference in the AUCs of two correlated ROC curves, assuming one AUC was 80% and the other was at least 82% and the prevalence of true matches lied within the range of 5% to 95%, commonly seen in record linkage applications. These two ROC curves were derived based on the two matching algorithms using combinations of the four enhancements. We assumed a moderate correlation of 50% between matching status determined by the two matching algorithms, recognizing that the power increases as the matching algorithms produces more correlated results.

Specific Aim 3: Measure the matching accuracy improvements resulting from using combinations of three best practice matching policy recommendations and four novel matching algorithm enhancements.

This aim seeked to evaluate which combinations of process and technical enhancements produce the greatest matching accuracy improvements. Our proposed work recognized that improving match accuracy requires both process and technical innovation. While matching accuracy improvements may be partially achieved by either independently enhancing matching data using three best-practice policy recommendations or enhancing matching methods using four novel algorithm innovations, we *hypothesized* that combining both best practice recommendations and algorithm innovations would maximize improvements in matching sensitivity, specificity, positive predictive value, and AUC. Conversely, we further *hypothesized* that not all combinations of process and technical enhancements would result in significant improvements in matching accuracy.

Justification. Because implementing process and technical innovations both have ongoing associated operational costs and potential privacy liabilities (e.g., by adding more identifying matching fields), it was important to develop evidence-based prioritizations for future matching enhancements. Consequently, evidence derived from assessing the relative improvements in matching accuracy among various combinations of both process and algorithm enhancements can inform future national matching strategy discussions.

Using the four gold standard patient matching datasets, we assessed the effectiveness of each remaining combination not evaluated in aims 1 and 2 by comparing enhanced matching results to baseline matching results derived from the original unmodified dataset and unmodified algorithm. To reduce the computational and analytical burden, we excluded any best practice recommendations and algorithm enhancement combinations from aims 1 and 2 that fail to show significant improvements in accuracy measures above baseline.

Research Design. We evaluated matching accuracy improvements resulting from implementing the three best-practice policy recommendations and the four novel matching algorithm innovations using sensitivity, specificity, PPV, and AUC. Using the four gold standard patient matching datasets, we assessed the effectiveness of combined process and algorithm enhancements by comparing enhanced matching results to baseline matching results derived from the original unmodified FS algorithm and unmodified gold-standard data sets. Table 2 highlights the specific analyses for Aim 3, where data enhancements ('D') and matching algorithm enhancements ('A') were combined. Each of the 105 matching enhancement combinations that exhibited significant accuracy improvements in prior aims were applied to the four gold standard matching datasets, for a total of up to 420 matching analysis datasets. We performed four accuracy analyses (sensitivity, specificity, PPV and AUC) per dataset for a total of up to 1,680 analyses.

Analysis. We measured the effect of combined process and technical matching enhancements on patient matching using sensitivity, specificity, PPV, and AUC. Sensitivity, specificity, and PPV were evaluated using proportions and 95% confidence intervals. To account for the clustering effect of multiple methods applied to the same record pair and assess the effects of individual and combined enhancements on matching performance, we performed a marginal logistic regression using generalized estimating equations[54,55]. For each metric we included main effects, two-way interactions, and three-way interactions of the enhancements in the model to allow differential effects of a factor as other factors vary. The standard errors of the accuracy measures were calculated using the robust sandwich variance estimation methods. The point estimate and the 95% confidence interval of the AUC, as well as the comparison of multiple AUCs, were performed using a nonparametric approach[56]. Comparison of these accuracy metrics between a) the unmodified FS algorithm and the four baseline gold standard data sets and b) the up to 105 matching enhancement combinations was performed using a multiple comparison approach with a Bonferroni adjustment.

Sample Size & Power. Sample sizes for the four gold standard record matching datasets ranged from 11,000 to 17,000. These sample sizes provided at least 80% power to detect a minimum of 2% difference in the AUCs of two correlated ROC curves, assuming one AUC is 80% and the other is at least 82% and the prevalence of true matches lies within the range of 5% to 95%, commonly seen in record linkage applications. These two ROC curves were derived based on the two matching algorithms using combinations of the four enhancements. We assumed a moderate correlation of 50% between matching status determined by the two matching algorithms, recognizing that the power increases as the matching algorithms produces more correlated results.

Results

Token Frequency

We found that incorporating token frequency information into the matching process can increase match accuracy, particularly when the quality of matching field data (e.g., data completeness, discriminating power) is suboptimal. Practitioners should consider incorporating field frequency into the matching process when correlating with incomplete or low-discriminating data.

Conditional Dependence

We examined different approaches to address conditional dependence (underlying correlation between agreeing fields). Accounting for dependency among the nonmatch class rather than the true match class resulted in increased accuracy. Although we hypothesized that accommodating dependency within both the match and nonmatch classes would produce superior performance, nonmatch record pairs provide a better signal for identifying correlation structures. One should consider including interaction terms for nonmatch class fields having a pseudo-R2 of 0.1% or greater.

Data Standardization

Data standardization exhibited modest match performance improvement. Data standardization eliminates spelling and formatting variations in demographic fields that may cause a missed match. For example, the last name of "O' Brien" would be standardized by uppercasing all letters and removing extra spaces to create "O'BRIEN." The most significant increase in match accuracy resulting from standardization was observed for the newborn data set, where data quality was much lower due to limited and changing matching identifiers (e.g., given and surname, lack of SSN) around the time of the patient's birth. Data standardization should be considered for data sets containing significant spelling and typographical variation to maximize match accuracy.

Similarity Index

Incorporating string similarity measures was associated with increased match accuracy. String similarity measures compare two strings (e.g., a name, an address) to determine how closely they agree. These measures typically produce a value from 0 to 1, where 0 represents no agreement, and 1 represents exact agreement between strings. We found that a dichotomized measure with a single threshold increased optimal accuracy. Where feasible, we encourage the use of string comparators.

Missingness

We evaluated three methods for accommodating missingness: treating missing data as MAR, MNAR, and MAD. MAR exhibited a significant match accuracy increase. In the general case, when demographic data are missing, the corresponding fields are treated as disagreeing, which results in a lower match score and is more likely not to match. The MAR method uses statistical formulas to impute a probable value for the field, which enables the field to contribute a partial score to the matching algorithm. We strongly recommend incorporating methods for accommodating missing data into match processes needing improved performance where practical and feasible.

Token Selection

We applied a machine learning method called "decision trees" (specifically an XGBoost decision tree) to determine the combination of token fields that would maximize match accuracy. Applying this method to a large, heterogeneous clinical data source produced performance metrics (PPV, sensitivity, specificity, and F score) above 97%, which indicates that such models can perform well when used with high-quality matching data. Additional performance gains may be achieved by applying a more complex XGBoost classification model, and we recommend that, though complex, machine learning models may significantly match accuracy and should be explored further.

Generalizability

The INPC represents a unique in vivo laboratory to evaluate real-world patient matching methods, and the demographics of the HIE catchment area closely mirror many of the demographics of the United States overall, supporting the generalizability of findings. Fields used for linkage are commonly available and include SSN, name (last, first, middle initial), sex at birth, date of birth (day, month, and year), telephone number, street address, and ZIP Code.

Limitations

Although the data included in this study represent a broad spectrum of healthcare settings, which supports the likelihood of generalizability, our analysis used data specific to Indiana health systems. Consequently, results may vary in environments with markedly differing demographic data characteristics and differing availability of matching fields. For example, the matching approaches described may produce more false-positive matches over a larger or different population.

Future Research

We seek to understand the performance of matching methods among various racial, ethnic, and other demographic subgroups. Increasing attention and concern are directed at algorithmic bias, and matching methods are no less susceptible to potential biases. Understanding how match algorithms perform among different subgroups is essential to ensuring high-quality research data and strengthening consumer trust in the validity of scientific results derived from data integrated via linkage.

Conclusions

Ensuring accurate and robust methods for integrating data from several sources for research purposes is essential to high-quality innovation and discovery. Through this research we have identified opportunities for improving linkage, including pursuing probabilistic (vs heuristic, or rule-based) matching methods and incorporating methods for accommodating missingness. Standardization and frequency-based approaches may help improve overall matching when working with lower-quality data. Finally, more work is needed to determine the performance of these methods among various subgroups, including racial and ethnic minorities.

List of Publications and Products

1. Kasthurirathne SN, Grannis SJ. Machine Learning Approaches to Identify Nicknames from A Statewide Health Information Exchange. AMIA Jt Summits Transl Sci Proc. 2019;2019:639-47.
2. McNutt AT, Grannis SJ, Bo N, Xu H, Kasthurirathne SN. Comparison of Supervised Machine Learning and Probabilistic Approaches for Record Linkage. AMIA Informatics summit 2019 Conference Proceedings.
3. Grannis SJ, Xu H, Vest JR, Kasthurirathne S, Bo N, Moscovitch B, Torkzadeh R, Rising J. Evaluating the effect of data standardization and validation on patient matching accuracy. Journal of the American Medical Informatics Association. 2019 May;26(5):447-56.
4. Grannis S, Kasthurirathne S, Bo N, Huiping X, editors. Evaluating Two Approaches for Parameterizing the Fellegi-Sunter Patient Matching Algorithm to Optimize Accuracy2019: Medinfo conference proceedings.
5. Xu H, Li X, Shen C, Hui SL, Grannis S. Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? The Annals of Applied Statistics. 2019;13(3):1753-90, 38.
6. Grannis SJ, Li X, Xu H, Ong TC, Kahn MG, Lembcke LR, et al., editors. Novel Application of Data Quality Metrics to Tailor Standardization of Patient Matching Fields. AMIA; 2020.
7. Grannis SJ, Kho AN, Phua J, Kasthurirathne SN. Evaluation of Token Collections and Matching Models to Support Privacy-Preserving Record Linkage (PPRL). InAMIA 2021.
8. Xu H, Li X, Grannis S. A simple two-step procedure using the Fellegi-Sunter model for frequency-based record linkage. J Appl Stat. 2022;49(11):2789-804.

9. Li X, Xu H, Grannis S. The Data-Adaptive Fellegi-Sunter Model for Probabilistic Record Linkage: Algorithm Development and Validation for Incorporating Missing Data and Field Selection. J Med Internet Res. 2022;24(9):e33775.
10. Williams KS, Grannis SJ. Patient-Centered Data Home: A Path Towards National Interoperability. Frontiers in Digital Health. 2022 Jul 13;4:887015.
11. Grannis SJ, Williams JL, Kasthuri S, Murray M, Xu H. Evaluation of real-world referential and probabilistic patient matching to advance patient identification strategy. Journal of the American Medical Informatics Association. 2022 Aug 1;29(8):1409-15.
12. Kiernan D, Carton T, Toh S, Phua J, Zirkle M, Louzao D, Haynes K, Weiner M, Angulo F, Bailey C, Bian J. Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet®, the National Patient-Centered Clinical Research Network. BMC Research Notes. 2022 Oct 31;15(1):337.
13. Gupta AK, Kasthurirathne SN, Xu H, Li X, Ruppert MM, Harle CA, Grannis SJ. A framework for a consistent and reproducible evaluation of manual review for patient matching algorithms. Journal of the American Medical Informatics Association. 2022 Dec 1;29(12):2105-9.
14. Marsolo K, Kiernan D, Toh S, Phua J, Louzao D, Haynes K, Weiner M, Angulo F, Bailey C, Bian J, Fort D. Assessing the impact of privacy-preserving record linkage on record overlap and patient demographic and clinical characteristics in PCORnet®, the National Patient-Centered Clinical Research Network. Journal of the American Medical Informatics Association. 2023 Mar 1;30(3):447-55.

Bibliography

1. McDonald CJ, Overhage JM, Dexter PR, et al. Canopy computing using the web in clinical practice. JAMA 1998; 280:1325-1329. PMID: 9794311
2. Finnell JT, Overhage JM, Grannis SJ. All Health Care is Not Local: An Evaluation of the Distribution of Emergency Department Care Delivered in Indiana. AMIA Annu Symp Proc. 2011:409-16. PMCID: PMC3243262
3. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010 Nov 10;2(57):57cm29. PMID: 21068440
4. Mason AR, Barton AJ. The emergence of a learning health care system. Clin Nurse Spec. 2013 Jan-Feb;27(1):7-9. PMID: 23222020
5. Appavu SI. Unique patient identifiers: What are the options? Journal of AHIMA, 1999;70(9) 50–57. PMID: 10977406
6. Hillestad R. Identity Crisis? Approaches to Patient Identification in a National Health Information Network. 2008. Retrieved April 19, 2014 from http://www.rand.org/content/dam/rand/pubs/research_briefs/2008/RAND_RB9393.pdf
7. Shekelle PG, Morton SC, Keeler EB. Costs and Benefits of Health Information Technology. Evidence Report/Technology Assessment No. 132. (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-02-0003.) AHRQ Publication No.06-E006. Rockville, MD: Agency for Healthcare Research and Quality. 2006. PMID: 17627328
8. Quantin C, Allaert FA, Avillach P, et al. Building application-related patient identifiers: what solution for a European country? Int J Telemed Appl. 2008:678302. PMCID: PMC2288643
9. Jain AK, Flynn P, Ross AA. Handbook of Biometrics. Springer, 2008.
10. Prabhakar S, Pankanti S, Jain AK. Biometric Recognition: Security and Privacy concerns. IEEE Security & Privacy. 2003 (3)2:33-42.
11. Gill L. Methods for automatic record matching and linking and their use in national statistics. Newport, UK: Office for National Statistics. 2001.
12. Lui S. Development of record linkage of hospital discharge data for the study of neonatal readmission. Chronic Dis Can. 1999;20(2):77-81. PMID: 10455039
13. Knuth DE. The Art of Computer Programming, Volume 3/Sorting and Searching, Second Edition. Addison-Wesley Publishing Company, 1998.
14. Lynch BT, Arends WL. Selection of a surname coding procedure for the SRS record linkage system. Washington, DC: U.S. Department of Agriculture, Sample Survey Research Branch, Research Division. 1977.
15. Verykios VS, Moustakides GV, Elfeky MG. A Bayesian decision model for cost optimal record matching. The VLDB Journal. 2003;12(1):28-40.
16. Borthwick A. A maximum entropy approach to named entity recognition. (Doctoral dissertation). New York University. New York, NY. 1999. ISBN:0-599-47232-4
17. Fellegi IP, Sunter SB. A theory of record linkage. Journal of the American Statistical Association, 1969;64(328):1183–1210.
18. Xu H, Hui SL, Grannis S. Optimal two-phase sampling design for comparing accuracies of two binary classification rules. Stat Med. 2014 Feb 10;33(3):500-13. PMID: 24038175
19. Grannis SJ, Banger A, Harris D. Perspectives on Patient Matching: Approaches, Findings, and Challenges. Agency for Health care Research and Quality and Office of the National Coordinator for Health Information Technology. Retrieved April 28, 2014 from https://www.healthit.gov/sites/default/files/patient-matching-white-paper-final-2.pdf.
20. Consistent Nationwide Patient Data Matching Strategy. Health Information Management System Society 2013 Policy Summit Congressional Ask #1 Recommendation to Congress. Retrieved April 2,

2014 from https://www.himss.org/files/HIMSSorg/Congressional_Asks_1%202013_InteroperabilityPatientID.pdf. 2013, September.

21. Marchibroda J. Challenges and strategies for accurately matching patients to their health data. Bipartisan policy Center. Retrieved Jan 2, 2014 from https://bipartisanpolicy.org 2012, June.
22. Health IT: Setting The Foundation To Transform Our Future. eHealth Initiative Government Affairs Retreat. Retrieved April 20, 2014 from https://www.ehidc.org/resources/ehi-government-affairs-retreat-2014-health-it-setting-foundation-transform-our-futurehttps://www.ehidc.org/sites/default/files/resources/files/2014-03-11-eHI-2014_Government_Affairs_Retreat-Summary_0.pdf. 2014, February.
23. Linking Health Care Information: Proposed Methods for Improving Care and Protecting Privacy. The Markle Foundation. Retrieved March 20, 2014 from http://www.markle.org/publications/863-linking-health-care-information-proposed-methods-improving-care-and-protecting-priv. 2005, February.
24. Morris G. Patient identification and matching initial findings. HealthIT.gov: the official site for Health IT information. Retrieved Jan 22, 2014 from http://www.healthit.gov. 2013, December 16.
25. Tang P. Recommendations to the Department of Health and Human Services on patient matching. Health IT Policy Committee: Recommendations to the National Coordinator for Health IT. Retrieved March 3, 2014 from http://www.healthit.gov. 2011, February 8.
26. U.S. Government Accountability Office. (2014, March). Electronic Health Records: HHS Strategy to Address Information Exchange Challenges Lacks Specific Prioritized Actions and Milestones. (Publication No. GAO-14-242). Retrieved April 2, 2014 from http://www.gao.gov/assets/670/661846.pdf.
27. 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. PMCID: PMC2744724
28. Daggy J, Xu H, Hui S, Grannis S. Evaluating Latent Class Models with Conditional Dependence in Record Linkage. Statistics in Medicine. 2014 Oct 30;33(24):4250-65. Doi: 10.1002/sim.6230. PMID: 24935712
29. Rock MJ, Levy H, Zaleski C, Farrell PM. Factors accounting for a missed diagnosis of cystic fibrosis after newborn screening. Pediatr Pulmonol. 2011 Dec;46(12):1166-74. doi: 10.1002/ppul.21509. PMCID:PMC4469987
30. Hoff T, Ayoob M, Therrell BL. Long-term Follow-up Data Collection and Use in State Newborn Screening Programs. *Arch Pediatr Adolesc Med.* 2007;161(10):994-1000. doi:10.1001/archpedi.161.10.994. PMID: 17909144
31. Grannis S, Biondich P, Downs S et al. Leveraging open-source matching tools and health information exchange to improve newborn screening follow-up. Public Health Information Network Annu Symp. Progress 2008.
32. Wu J, Xu H, Finnell JT, Grannis SJ. A Practical Method for Predicting Frequent Use of Emergency Department Care Using Routinely Available Electronic Registration Data. AMIA Annu Symp Proc. 2013:1524.
33. 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. PMID: 15360771
34. Devore S, Champion RW. Driving population health through accountable care organizations. Health Aff (Millwood). 2011 Jan;30(1):41-50. doi: 10.1377/hlthaff.2010.0935. PMID: 21209436
35. Wu FM, Rundall TG, Shortell SM, Bloom JR. Using health information technology to manage a patient population in accountable care organizations. J Health Organ Manag. 2016 Jun 20;30(4):581-96. doi: 10.1108/JHOM-01-2015-0003. PMID: 27296880

36. McWilliams JM, Hatfield LA, Chernew ME, Landon BE, Schwartz AL. Early Performance of Accountable Care Organizations in Medicare. N Engl J Med. 2016 Jun 16;374(24):2357-66. doi: 10.1056/NEJMsa1600142. PMID: 27075832
37. Daggy JK, Xu H, Hui SL, Gamache RE, Grannis SJ. A practical approach for incorporating dependence among fields in probabilistic record linkage. BMC Med Inform Decis Mak. 2013 Aug 30;13:97. PMCID: PMC3766252
38. 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. PMCID: PMC1479910
39. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. Proc AMIA Symp. 2002:305-9. PMCID: PMC2244404
40. Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. National Academies Press, 1999.
41. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. BMJ. 2016 May 3;353:i2139. doi: 10.1136/bmj.i2139.
42. Joel Rodrigues (2010). Health Information Systems: Concepts, Methodologies, Tools, and Applications, Volume 1. IGI Global. ISBN 978-1-60566-988-5.
43. Council for Affordable Quality Health care. (2011). Normalizing Patient Last Name Rule. Retrieved Dec 18, 2013 from http://www.caqh.org. 2011, March.
44. Health Care Claim: Institutional. American National Standards Institute ASC X12 Standard 837, version 003070: 1997. Accessed March 7, 2014 from https://www.cms.gov/medicare/billing/electronicbillingeditrans/downloads/5010a2837acg.pdf
45. International Country Codes. International Standard ISO 3166-3: 2010. Accessed Feb 4, 2013 from http://www.iso.org/iso/country_codes.htm.
46. Mailing Standards of the United States Postal Service Publication 28 - Postal Addressing Standards. USPS Publication 28: 2013. PSN 7610-03-000-3688. Accessed February 20, 2014 from http://pe.usps.gov/cpim/ftp/pubs/pub28/pub28.pdf.
47. Series E: Overall Network Operation Telephone Service, Service Operation and Human Factors. ITU-T Recommendation E.123. Accessed May 27, 2016 from https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-E.123-200102-I!!PDF-E&type=items.
48. Winkler WE. Methods for evaluating and creating data quality. Information Systems 2004;29(7):531–550.
49. Herzog TN, Scheuren FJ, Winkler WE. (2007). Data Quality and Record Linkage Techniques. New York: Springer.
50. Dixon BE, Rosenman M, Xia Y, Grannis SJ. A vision for the systematic monitoring and improvement of the quality of electronic health data. Stud Health Technol Inform. 2013;192:884-8. PMID: 23920685
51. Dixon BE, McGowan JJ, Grannis SJ. Electronic laboratory data quality and the value of a health information exchange to support public health reporting processes. AMIA Annu Symp Proc. 2011;2011:322-30. PMCID: PMC3243173
52. Comer KF, Grannis S, Dixon BE, Bodenhamer DJ, Wiehe SE. Incorporating geospatial capacity within clinical data systems to address social determinants of health. Public Health Rep. 2011 Sep-Oct;126 Suppl 3:54-61. PMCID: PMC3150130
53. https://ribbs.usps.gov/index.cfm?page=address_quality. Accessed May 10, 2016.
54. Leisenring W, Alono T, Pepe MS. Comparisons of Predictive Values of Binary Medical Diagnostic Tests for Paired Designs. Biometrics. 2000 Jun;56(2):345–351. doi: 10.1111/j.0006-341X.2000.00345. PMID: 10877288
55. Leisenring W, Pepe MS, Longton G. A marginal regression modelling framework for evaluating medical diagnostic tests. Stat. Med., 16 (11) (1997), pp. 1263–1281 PMID: 9194271

56. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837–845. PMID: 3202132
57. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC Health Serv Res. 2010 Feb 18;10:41. doi: 10.1186/1472-6963-10-41.
58. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1977;39(1):1-38
59. Levenshtein V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 1966;10(8):707-710.
60. Sidelli R, Friedman C. Validating patient names in an integrated clinical information system. In: Symposium on Computer Applications in Medical Care; 1991; Washington, D.C.; 1991. p. 588-592. PMCID: PMC2247599
61. Corbett J, Gu CC, Rice JP, Reich T, Province MA, Rao DC. Power loss for linkage analysis due to the dichotomization of trichotomous phenotypes. Hum Hered. 2004;57(1):21-7. PMID: 15133309
62. Tsuruta H, Bax L. Polychotomization of continuous variables in regression models based on the overall C index. BMC Med Inform Decis Mak. 2006; 6: 41. PMCID: PMC1770908
63. Tromp M, Méray N, Ravelli ACJ, Reitsma JB, Bonsel GJ. Ignoring Dependency between Linking Variables and Its Impact on the Outcome of Probabilistic Record Linkage Studies. Journal of the American Medical Informatics Association: JAMIA. 2008;15(5):654-660. doi:10.1197/jamia.M2265. PMCID: PMC2528043
64. Wang Z., Zhou XH, Wang M. Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard. Biostatistics 2011; 12(3): 567-581. PMCID: PMC3114651
65. Tromp M, Ravelli ACJ, Meray M, Reitsma JB, Bonsel GJ. An efficient validation method of probabilistic record linkage including readmissions and twins. Methods of Information in Medicine 2008; 47(4): 356-363. PMID: 18690369
66. Enders CK, Fairchild AJ, MacKinnon DP. A Bayesian Approach for Estimating Mediation Effects with Missing Data. Multivariate Behav Res. May 1, 2013; 48(3): 340–369. PMCID: PMC3769802
67. Liu WZ, White AP, Thompson SG, Bramer SG. Techniques for dealing with missing values in classification. Liu X, Cohen P, Berthold (Eds.), Adv intell data anal reason data, Springer, Berlin, Heidelberg (1997), pp. 527–536.
68. Saar-Tsechansky, Provost MF. Handling missing values when applying classification models. J Mach Learn Res, 8 (2007), pp. 1625–1657
69. Tromp M, Reitsma JB, Ravelli AC, Méray N, Bonsel GJ. Record linkage: making the most out of errors in linking variables. AMIA Annu Symp Proc. 2006:779-83. PMCID: PMC1839331