MINI-SENTINEL DATA ACTIVITIES

ANONYMOUS LINKING OF DISTRIBUTED DATABASES

Prepared by: Alvin Mushlin, MD, ScM,[1] Curtis Baginski, BS,[3] Carlos Bell, MPH,[2] Bonnie Brown, BS, MS,[3] Robert Dickson, BS,[3] Susan Forrow, BA,[4] Chunfu Liu, ScD,[5] Stephen Lyman, PhD,[1] David McDermott, BS, MS,[3] Brian Macy, BS,[3] Tobi Moriarty, BS,[3] Tom Puenpatom, PhD,[5] Mitra Rocca, PhD,[2] Paul Romagano, BS,[3] Lucas Romero, MPA,[1] Art Sedrakyan, MD, PhD,[1] Nandini Selvam, PhD, MPH,[5] Jeremy Rassen, ScD[6]

Author Affiliations: 1. Weill Cornell Medical College, Department of Public Health, New York, NY. 2. Office of Medical Policy, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD. 3. IBM InfoSphere Solutions, Bethesda, MD. 4. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA. 5. HealthCore, Inc., Wilmington, DE. 6. Division of Pharmacoepidemiology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.

August 30, 2013

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel Data Activities

Anonymous Linking of Distributed Databases

Table of Contents

I. Introduction 1
A. Objectives 1
B. Overview 2
1. Initial Requirements 2
2. Selection of Software Vendor 2
3. Report Structure 3
C. Literature Review 3
D. Data Sources 4
1. Hospital for Special Surgery/Weill Cornell Medical College CERT/Legacy Cohort Data (Registry) 4
2. WellPoint/HealthCore, Inc. Data (Claims) 5
II. Matching Procedures 6
A. Gold Standard Linkage 6
1. IBM's Quality Stage Process 6
B. Anonymous Linkage 7
1. IBM's Anonymous Resolution Process 7
2. Further Detail on the Hashing and Matching Process 11
3. Anonymous Resolution Output 12
III. Accuracy Analysis 13
A. Overview 13
B. Statistical Considerations and Analysis 15
IV. Results 17
A. Phase 1 17
1. Phase 1 Results 17
2. False Negatives and False Positives 18
3. Summary of Findings and Recommendations 21
B. Phase 2 22
1. Terminology Clarifications 22
2. Configuration Changes 23
3. Phase 2 Results 24
4. Summary of Findings and Recommendations 27
V. Conclusions and Implications 28
VI. Literature Review Bibliography 30

I. INTRODUCTION

A. OBJECTIVES

Mini-Sentinel is a pilot project conducted under contract with the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, known as the Sentinel System, for monitoring the safety of FDA-regulated medical products.

The objectives of the Mini-Sentinel Anonymous Linking project were to explore the feasibility of linking multiple health care databases without the need to share information that directly identifies patients, to identify barriers to anonymous linkage within the Mini-Sentinel network, and to provide guidance on the value and potential future directions for anonymous linkage of health care databases for medical product safety surveillance. Linkage of health care databases may provide more robust cross-sectional or longitudinal patient profiles that would enhance medical product safety surveillance evaluations and improve access to information that would not be present in claims data or Electronic Health Records (EHRs) alone. Health care utilization information tracking the services received and the technology used to diagnose, treat or manage disease and illness is now largely available in electronic format through the claims submitted by healthcare providers to insurance companies or government payers. Additionally, information about individual patients, their health outcomes and the details of their care is present electronically in their medical records or in registries designed to collect information on diseases, procedures, or devices.

The project required participation by two Mini-Sentinel data partners with overlapping memberships and patient populations. One of the partners selected was the New York Hospital for Special Surgery (HSS)[1]/Weill Cornell Centers for Education & Research on Therapeutics (CERT)[2] which maintains a total joint replacement device registry. The other was HealthCore, Inc.,[3] which holds and manages a health insurance database containing administrative claims data for WellPoint, Inc.,[4] one of the Blue Cross and Blue Shield Association health insurance companies. The anonymous linking approach we investigated, IBM's Anonymous Resolution (AR) software,[5] if successful, will enable two holders of protected health information (PHI) to derive aggregate treatment and health outcome information about their overlapping populations without requiring either data holder to disclose PHI or proprietary data to each other, or any other party. IBM's anonymous resolution technique, which falls within the US Department of Health and Human Services (HHS) "Safe Harbor" guidelines for matching with de-identified data,[6]

---

[1] http://www.hss.edu/

[2] US Department of Health and Human Services (HHS), Agency for Healthcare Research and Quality (AHRQ), Centers for Education & Research on Therapeutics (CERT) http://www.certs.hhs.gov/

[3] http://www.healthcore.com/home/

[4] http://www.wellpoint.com/

[5] See: Swire, P. Research Report: Application of IBM Anonymous Resolution to the Health Care Sector. IBM Corporation 2006. Available at: http://www.peterswire.net/anon.resolution.whitepaper.pdf

[6] HHS Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (2012). Available at: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

greatly reduces the potential for inadvertent disclosure of PHI. The accuracy of the anonymous linking process was assessed by comparing the results obtained using the AR software against results obtained by an equivalent match that used identifiable, rather than anonymized, data. This "gold standard" match was obtained using IBM's Quality Stage (QS) software.[7]

B. OVERVIEW

1. Initial Requirements

We attempted to anonymously link individuals from the two separate databases using only their names, dates of birth, and addresses and to assess the feasibility and accuracy of doing so. This required the selection of state-of-the-art software for anonymous linkage, as well as additional software that could match individuals common to both datasets using available, non-anonymized identifying information to serve as the gold standard against which to compare the accuracy of the anonymous matching process. Given the sensitive nature of the identifiable information required for the gold standard match, legal approvals from the participating data partners were required. Due to institutional policies and Federal regulations concerning PHI, legal agreements were required among the data partners sharing PHI, as well as with any third parties participating in the project that would have access to PHI, such as IBM.

From a logistical perspective, the computer platform needed to be located so as to minimize the need for transfer of PHI between parties or sites and to maximize performance for what was anticipated to be a computer-intensive process. In selecting the platform, we needed to balance the requirements for computing power with the budget for this project and potential future Mini-Sentinel or Sentinel implementations.

2. Selection of Software Vendor

To implement anonymous linkage, the workgroup considered a "buy vs. build" approach. In a "build" approach, software to facilitate the gold standard and/or anonymous matching would be created from the ground up. Due to the complexity and time required, this option was quickly deemed infeasible. In the alternative "buy" approach, the workgroup would use off-the-shelf software for creating the gold standard match and performing anonymous linkage. The creation of the gold standard match was considered a one-time process needed for a proof of concept that would not likely be a part of the Sentinel program in the future, while the anonymous linkage process was considered to be something that, if successful, could be replicated in future Mini-Sentinel and Sentinel projects.

The workgroup identified IBM's QS software as robust, proven software for performing the gold standard match. For the anonymous linkage, the workgroup, in consultation with health informatics experts in the group and at the FDA, conducted in-depth interviews with several vendors, including IBM, Oracle, and other firms. From these interviews, it was apparent that IBM's AR software was the only product currently in the marketplace that could offer the functionality required for this project.

---

[7] IBM Datasheet: IBM Infosphere Quality Stage Investigate, cleanse and manage high-quality data to deliver better business results. IBM Corporation 2011. Available at: http://public.dhe.ibm.com/common/ssi/ecm/en/imd11784usen/IMD11784USEN.PDF

Following an extensive presentation to the workgroup in which they displayed both their QS and AR technologies, IBM was chosen as the commercial vendor for this project because their products were considered sufficiently robust and promising to justify a formal evaluation of their feasibility and accuracy. IBM agreed to provide the expertise and software need for this project without charge.

3. Report Structure

In the following sections, we first briefly review the literature about prior attempts to achieve anonymous linkage of two different datasets and then describe in more detail the datasets chosen for this project. Next, we outline how a fully identified gold standard matching of individuals in both the device registry and the claims dataset was accomplished. Then we describe the procedures followed to accomplish the anonymous linkage and measure its effectiveness. Finally, we present the results obtained and discuss the implications going forward.

C. LITERATURE REVIEW

Using the Medline database, we sought to identify previous instances where anonymous linkage had been carried out successfully in a medical context. We did both structured and unstructured searches. For the structured search, we downloaded abstracts for all articles that were labeled with the "Confidentiality" and "Medical Record Linkage" MeSH terms. We reviewed all 274 articles and selected those that were relevant to our inquiry. For our unstructured search, we examined the reference lists of the relevant articles to determine whether there was additional literature to be included; we added several articles to our bibliography as a result.

Like many terms, "anonymous linkage" appears to mean different things to different research groups. Some of the literature we examined claimed that the methods included anonymous linkage, but the approach would not meet the needs of Mini-Sentinel. Methods used included anonymization by aggregation (looking at groups of patients rather than individuals), use of identifying numbers rather than names, and use of randomly-system-generated identifiers (such as study ID). We have included these articles in the bibliography but omitted them from our summary findings.

The remainder of the articles we examined all utilized secure hash algorithm (SHA) based exact matching approaches. Specifically, they encrypted linking identifiers from each of two sources, and then established links based on whether the encrypted identifiers matched. As an example, consider the two possible linkages shown in Table 1. In each case, it is possible to tell whether the records match by looking only at the encrypted information; no unencrypted identifiers are required.

Table 1. Example of SHA encryption-based linkage approach

Source A	Source A (Encrypted)	Source B	Source B (Encrypted)	Match?
John Smith	KRV2h+l0}Mo#Ao	John Smith	KRV2h+l0}Mo#Ao	Yes
Jane Smythe	G&FNd=4y.hDo5z	Richard Smithward	r1oBUKY?8t+E(a	No

Many of the papers reported the specificity and sensitivity of the specific anonymous linkage that they attempted. Specificities ranged from single digits to 100%, while sensitivities ranged from single digits to 95%.

A large number of the articles we examined originated from France. Based on the background sections of these works, it appears that French and European law has compelled attention to these aspects of patient privacy for some time (from at least the mid-1990s), and that the laws in France may be stricter than the Health Insurance Portability and Accountability Act (HIPAA)[8] standard in the United States.

Our conclusion from the literature search is that although multiple attempts have been made to link healthcare data sources in an anonymous fashion and results to-date have been encouraging, none has been studied with sufficient rigor to obviate the need for, and importance of, this project . The lack of a validated method for anonymous linkage of health care data across disparate data sources reinforced the value of this effort by Mini-Sentinel to assess the feasibility and accuracy of an anonymous linking process against a gold standard. The findings enabled FDA to assess the potential for augmenting the data sources available for the Agency's post-marketing safety surveillance efforts.

D. DATA SOURCES

The workgroup selected two Mini-Sentinel data partners that were likely to have overlapping patient populations and met the characteristics of an anonymous linkage scenario pertinent to Mini-Sentinel, which would involve linking administrative claims information from a Mini-Sentinel data partner to additional information held by disease or device registries or other comparable sources. First, we obtained data from a registry of orthopedic procedures carried out at the Hospital for Special Surgery. We sought to link these patients with claims data from a large commercial health plan, WellPoint, using data held by their subsidiary, HealthCore, Inc. Patients common to both data sources and available for linkage would have been treated at the HSS and covered by a WellPoint insurance plan. If all available data were matched, the resulting dataset would merge the longitudinal health history represented by the claims data with the detailed information about the orthopedic procedures and patients' self-reported functional status information stored in the registry. Because information about health conditions and services received was not necessary for this project, we only used names, dates of birth, and addresses.

1. Hospital for Special Surgery/Weill Cornell Medical College CERT/Legacy Cohort Data (Registry)

Founded in 1863, HSS is the nation's oldest orthopedic hospital. More than 25,000 surgical procedures are performed annually. HSS performs more hip surgeries and more knee replacements than any other hospital in the nation and is nationally ranked #1 in orthopedics, #3 in rheumatology, and #10 in neurology by *U.S. News & World Report* (2012-2013). HSS has been top-ranked in the Northeast for both orthopedics and rheumatology for the 22nd consecutive year. In addition, *Consumer Reports* ranked HSS as the best hospital in New York City according to their patient satisfaction study.

[8] http://www.hhs.gov/ocr/privacy/index.html

The CERT program is part of a national research initiative, funded by the federal Agency for Healthcare Research and Quality (AHRQ), to optimize the safety and effectiveness of medical care in the United States. There are currently over 200 completed and ongoing studies sponsored by the CERT program aimed at improving the health of all Americans.

HSS partnered with the Weill Cornell Medical College (WCMC) CERT program to study the outcomes in patients who have their knee, hip or shoulder replaced. The WCMC-HSS CERT registry[9] focuses exclusively on total joint replacements, and is among the most comprehensive registries of its kind in the country.

The HSS Hip and Knee Replacement CERT/Legacy Cohort contains data from all patients who had hip or knee replacement surgery from May 1, 2007 to January 31, 2011. Two sets of data for each individual patient were collected. The first set was from the patients themselves who were scheduled to have hip or knee replacement surgery. Patients were consented and given a survey pre-operatively and at 6 months, 2 years and 5 years after surgery. The baseline survey consisted of questions on function, pain, activity, quality of life, expectations, and previous joint replacement using standard instruments (SF-36, HOOS/KOOS, WOMAC, EuroQOL). Six month surveys measured adverse events (e.g., deep vein thrombosis, pulmonary embolus, myocardial infarction, and stroke). Those who completed the baseline survey were contacted for follow-up information at 2 and 5 years after surgery using the same standard instruments completed at baseline except for the expectations survey which was replaced with a satisfaction survey. Surveys were administered by mail, email, and in some cases, by telephone. The second set of data was collected from hospital administrative systems and includes demographics and procedure data (e.g., age, sex, race, and date of surgery, type of surgery, comorbidities, and types of devices/implants used). These data were collected electronically on all eligible patients regardless of their consent status.

The HSS registry dataset used for this project (consisting only of patients' names, dates of birth, and addresses) contained approximately 20,000 records before matching.

2. WellPoint/HealthCore, Inc. Data (Claims)

HealthCore, Inc., established 1996, is an independently operating, wholly owned subsidiary of WellPoint, Inc. WellPoint is among the largest health benefits company in terms of medical enrollment in the United States. HealthCore specializes in health outcomes and epidemiologic research, as well as drug, vaccine, and biologic safety evaluations. HealthCore provides research services to a variety of clients across the healthcare setting.

The data environment at HealthCore provides one of the largest commercial insurance research data environments in the United States. The HealthCore Integrated Research Database (HIRD^{SM})[10] is owned and operated by HealthCore and includes automated computerized claims data and enrollment information from 14 Blue Cross and/or Blue Shields (BCBS) licensed plans. As of May 2013, the HIRD contains data from approximately 46.5 million lives with medical coverage and 30.2 million lives with both medical and pharmacy coverage at any point from January 2006 through December 2012. The

[9] http://weill.cornell.edu/cert/pdf/joint_registry_brochure.pdf
[10] http://healthcore.com/home/research_enviro.php?page=Research%20Environment

HIRD contains fully adjudicated paid claims, with dates of service for all non-capitated ambulatory, emergency department, inpatient, and outpatient encounters. The HIRD also contains diagnostic laboratory testing results from two large national laboratories for WellPoint-affiliated health plan members receiving outpatient laboratory services. Data include full ranges of hematologic, chemistry, immunologic, and microbiologic (including culture and antibiotic sensitivity results) from over 10 million members. HealthCore also has the ability to link its data with various federal and state registries, and since the HIRD consists of completely identifiable information, it can be supplemented with medical records for a large proportion of the population.

The dataset extracted from HIRD for the Mini-Sentinel Anonymous Linkage project was for the study period between January 1, 2006 and January 31, 2012. Data were from health plans of Georgia, California, Virginia, New York, Nevada, Indiana, Kentucky, Missouri, Ohio, Wisconsin, Connecticut, Maine, and New Hampshire for members of fully insured and administrative services only (ASO)[11] opt-in status. All individuals having at least one day of medical eligibility during the study period were included. Members with multiple names and addresses had multiple records in the dataset.

The HealthCore dataset used for this project (consisting only of patients' names, dates of birth, and addresses) is labeled HC in this report; it contained approximately 43 million records before matching.

II. MATCHING PROCEDURES

A. GOLD STANDARD LINKAGE

1. IBM's Quality Stage Process

The gold standard match was carried out using IBM's Quality Stage (QS) software. QS is part of IBM's InfoSphere Information Server data integration platform. The core capabilities of the software include data investigation, data standardization, address verification, probabilistic matching, data survivorship, and data enrichment. QS was developed to help a variety of users match names and addresses with related data such as phone numbers, birth dates, email addresses, and other descriptive information, using ordinary, unencrypted text from two separate data sources. It relies on highly accurate probabilistic matching algorithms to match data elements which may vary slightly, such as "Ave" and "Avenue" or "John" and "Johnny". The probabilistic matching process used by QS creates a higher likelihood that a complete match will be accomplished, which means that the QS match can serve as the gold standard including all individuals common to both datasets.

QS consists of several steps:

- **Step 1 - Standardization:** The QS program cleanses the input data to standardize the data fields before matching. For example, if one of the matching fields is the address, the initial phase of QS will standardize all street names for consistent naming ("Street" rather than "St" or "St.", "9th"

[11] Administrative services only (ASO) populations are those individuals who belong to self-insured entities for whom the insurance company provides administrative services only.

rather than “Ninth”, “SW” rather than “Southwest”, and so forth). It uses built-in and user-specified rules, applied equally to both datasets, to minimize “noise” in the data while retaining meaningful variation (“9th Street” versus “9th Avenue”).

- **Step 2 – Matching:** The QS matching process detects duplicate records, inconsistencies, and missing values, and searches for matches between the two data sources. A built-in scoring process assigns a score to each match. A higher score indicates a better quality of match, and thus a higher probability that the match is true. An example of this process is shown in Figure 1.

- **Step 3 – Creating Linking Keys:** Once a match has been made, QS constructs linking keys so that the match can be uniquely identified. All processing is carried out in a scalable and parallelizable framework. Consequently, QS was projected to be sufficiently powered to handle the 43 million HealthCore records for this project on a computer server with reasonable capacity to process the data.

Figure 1. Example of Data Transformation

Classic transformation: account to customer

Account view

Source	Legacy Key	Name	Address	Phone	Birth Date	Cust-ID
Life	70328574	John Smith Jr.	10 Main St Boston MA 02110	781-259-9945	02/05/1940	
Home	80328575	Mr. John Smith	10 Main St Unit 10 Boston MA 02111	617-259-9000		
Auto	90238495	J. Smyth	Main St Bostan Mass 02110	781-295-9945	02/05/1941	

Customer view — ***Link related records to create cross-reference IDs***

Source	Legacy Key	Name	Address	Phone	Birth Date	Cust-ID
Life	70328574	John Smith Jr.	10 Main St Boston MA 02110	781-259-9945	02/05/1940	0001
Home	80328575	Mr. John Smith	10 Main St Unit 10 Boston MA 02111	617-259-9000		0001
Auto	90238495	J. Smyth	Main St Bostan Mass 02110	781-295-9945	02/05/1941	0002

Customer profile — ***Create a customer profile with the best information from all sources***

Source	Legacy Key	Name	Address	Phone	Birth Date	Cust-ID
CP		Mr. John Smith Jr.	10 Main St Unit 10 Boston MA 02111	617-259-9000	02/05/1940	0001
CP		J. Smyth	Main St Bostan Mass 02110	781-295-9945	02/05/1941	0002

B. ANONYMOUS LINKAGE

1. IBM’s Anonymous Resolution Process

The anonymous linkage aspect of the project was carried out using IBM’s AR software. Like QS, the AR software is part of IBM’s InfoSphere product line. AR works in a similar manner to the SHA method described in the literature review and displayed in Table 1. Further detail is provided below. In this approach, data is anonymized prior to leaving the originating source data site, and anonymized data is

matched deterministically on encrypted values. However, unlike standard SHA approaches, AR accommodates data variation such as nicknames, date-of-birth formats, and other inconsistencies in order to reduce variability in the input data and thus facilitate higher-quality deterministic matching. This is accomplished with three methods during data pre-processing: Expanded Hash generation, Address Standardization, and Address Validation.

- **Expanded Hash Generation:** Many data sources contain variations of common identifiers such as name and date of birth (DOB). Where one database might contain "Jonathan Smith", the next might contain "Johnny Smith". AR handles variation in these data types by generating additional hash values. For example, it will generate the root-name "Robert" for records with "Bobby". These additional hash values are used during the linkage step to appropriately match records.

- **Address Standardization:** Address components are often represented in different ways, for example, "123 Main St." and "123 Main Street". It is important to standardize the addresses to all follow the same format (123 Main Street) for optimal matching after the hashes are produced.

- **Address Validation:** Addresses can also contain erroneous information. For example "123 Main Parkway Street, Boulder Colorado" might be a data-entry error. Addresses can be validated against lists of valid addresses. This can enhance the data, correct errors, and single out invalid addresses.

Due to the matching of hash values, any small change in input value will yield a large change in the encrypted value, so cleansing and standardization of input data is critical. For the purposes of this project, where AR matches were to be compared to gold standard QS matches, *equivalent* approaches to cleansing of the input datasets were also important, so that AR matches had the best chance of equaling QS matches.

Similar to QS, AR works through a series of steps:

- **Step 1 - Pre-processor:** The AR pre-processor readies data in each dataset for the matching process, largely in support of standardizing the input data so that attempts to match hashed values across disparate datasets have the maximal chance to succeed. Specifically, the pre-processor will perform the name standardization and address validation and correction described above, as well as a normalization process that applies rules to phone numbers, dates of birth, social security numbers, and other significant attributes in preparation for hashing.

- **Step 2 - Anonymizer:** The AR anonymizer processes data through a one-way hash function, which applies a cryptographic hash to de-identify the data by transforming the values into a form that is computationally and mathematically irreversible. This is the process displayed in Table 1, where the inputs to the anonymizer are identifiable information elements such as name or address, and the outputs are encrypted values that cannot be transformed back to their original data. Further, with the expanded hash generation, the anonymizer will also hash *variants* of data elements in order to maximize matching success.

- **Step 3 – Resolver (Matcher):** The AR resolver receives the anonymized (hashed) data from multiple sources and performs linkage by identifying matches in the anonymized content database. This resolution takes place in a separate location, outside the site of the original data; the fully identified and plain text information is neither available nor required.

The one-way hashing process is critical to maintain the anonymity of the matches. A weak hash may allow a malicious individual to "reverse engineer" the hashed values into the original patient information. For that reason, a strong and mathematically proven hashing function is needed. AR supplies several defense-grade hashing algorithms, and augments the hashing process with a "salt" value, a randomly-generated string that adds a second layer of protection. In particular, the salt disallows the hashing of a dictionary of items (e.g., names) which would allow for easy reverse engineering.[12]

The steps of the AR process are shown in Figure 2 and Figure 3.

[12] If a hashing algorithm is unsalted, then anybody with access to the hashing algorithm could create a "dictionary" of values by hashing common names. A dictionary entry would show, for example, that the name "John Smith" maps to a hashed value of KRV2h+l0}Mo#Ao. Looking at encrypted data could enable those with a dictionary to reverse engineer the encrypted data back to the protected input values. Adding a so-called "salt" adds a layer of protection by requiring that someone know both the hashing algorithm and the particular salt value used to generate hashed values, and strongly inhibits the creation of a dictionary.

Figure 2. Anonymization Process 1

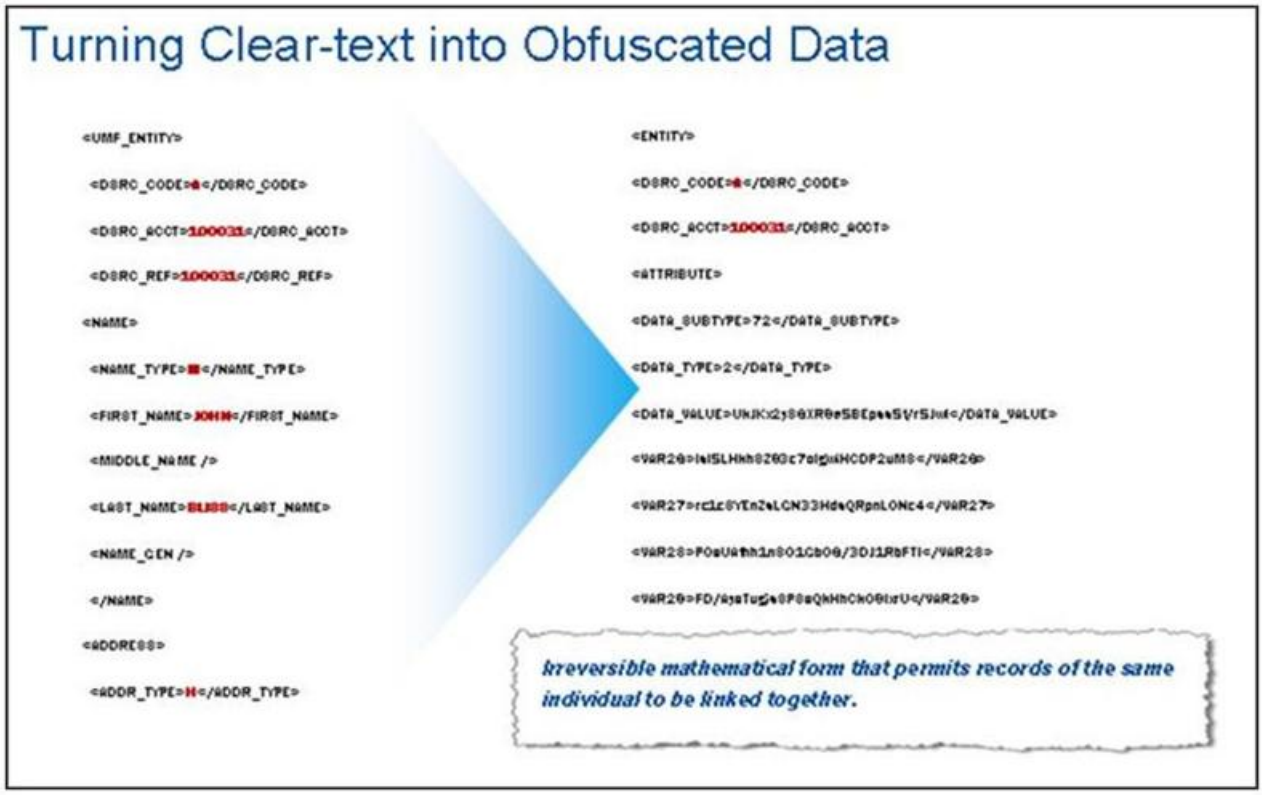
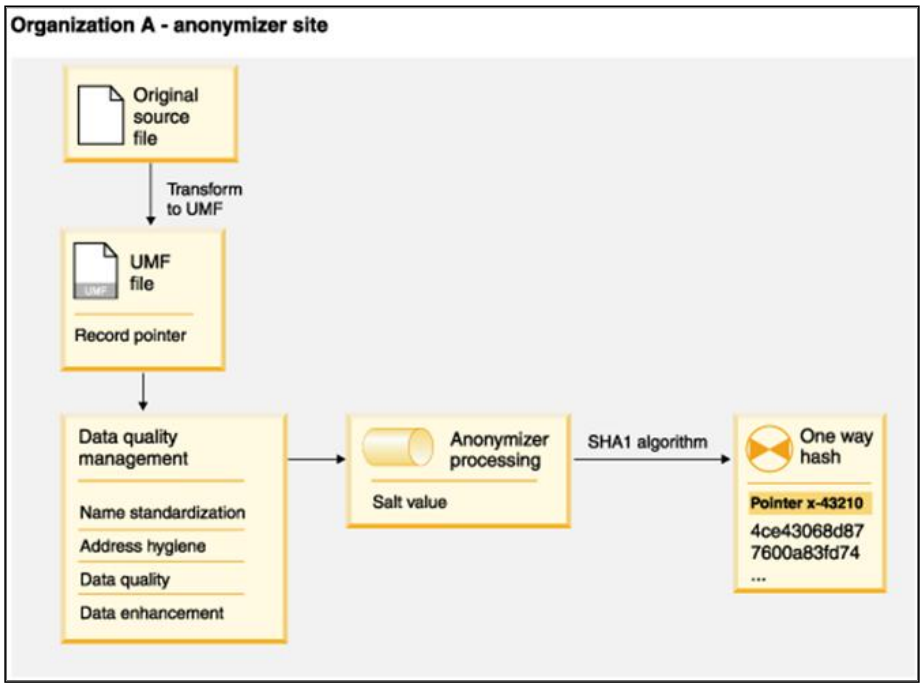


Figure 3. Anonymization Process 2



2. Further Detail on the Hashing and Matching Process

In general, anonymous linkage works by encrypting key data in both systems and then comparing the encrypted keys to see whether a match has occurred. Consider the example of two *likely true* matches carried out in the anonymous setting as shown in Table 2. Each system encrypts the full name and a match is made on the identical encrypted keys. It is important to note that even a small change in input data can cause a large change in encrypted values, causing a non-match. A typical SHA-encryption linkage approach does not handle variations such as nicknames, address variations, date of birth transpositions, or other inconsistencies.

Table 2. Typical SHA-encryption Linking Encryption

Source A	Source A (Encrypted)	Source B	Source B (Encrypted)	Match?
Robert Blake	Lw4V2h+l0}Mo%sfd	Robert Blake	Lw4V2h+l0}Mo%sfd	Yes
Robert Blake	Lw4V2h+l0}Mo%sfd	Bobby Blake	Tslwi45hslLLhk#50a	No

Unlike typical SHA-encryption approaches, IBM's AR software does handle certain variations in the data. Specific pre-processing steps improve the data quality, account for some typographical variations such as transposed dates, and expand encrypted keys to include name variations. To handle possible variations, IBM AR creates encrypted values for the exact input data *as well as likely variants.*

Consider the example shown in Table 3. Assume that Source A contains the record "John Smith", which encrypts to the value "KRV2h+l0}Mo#Ao ". Source B contains the record "Johnny Smith" which encrypts to "G&FNd=4y.hDo5z ". A traditional SHA-encryption scheme will not match these variations. AR will create an additional hash of the canonical name "Jonathan Smith" in each case to contribute to the resolution process. With the input data of "John Smith" in Source B, AR will create a canonical hash for that record as well, or "Jonathan Smith". Therefore at the Resolution site, AR will see both source systems contributed a "Jonathan Smith" and match them back to the original source records. While this example focuses on the name field, AR also creates additional encrypted records for address, date, and other fields.

Table 3. IBM AR Anonymous Linking Match Example

Source A	Source A (Encrypted)	Source B	Source B (Encrypted)	Match to Source B?
John Smith	KRV2h+l0}Mo#Ao	Johnny Smith	G&FNd=4y.hDo5z	Traditional SHA-encryption: No

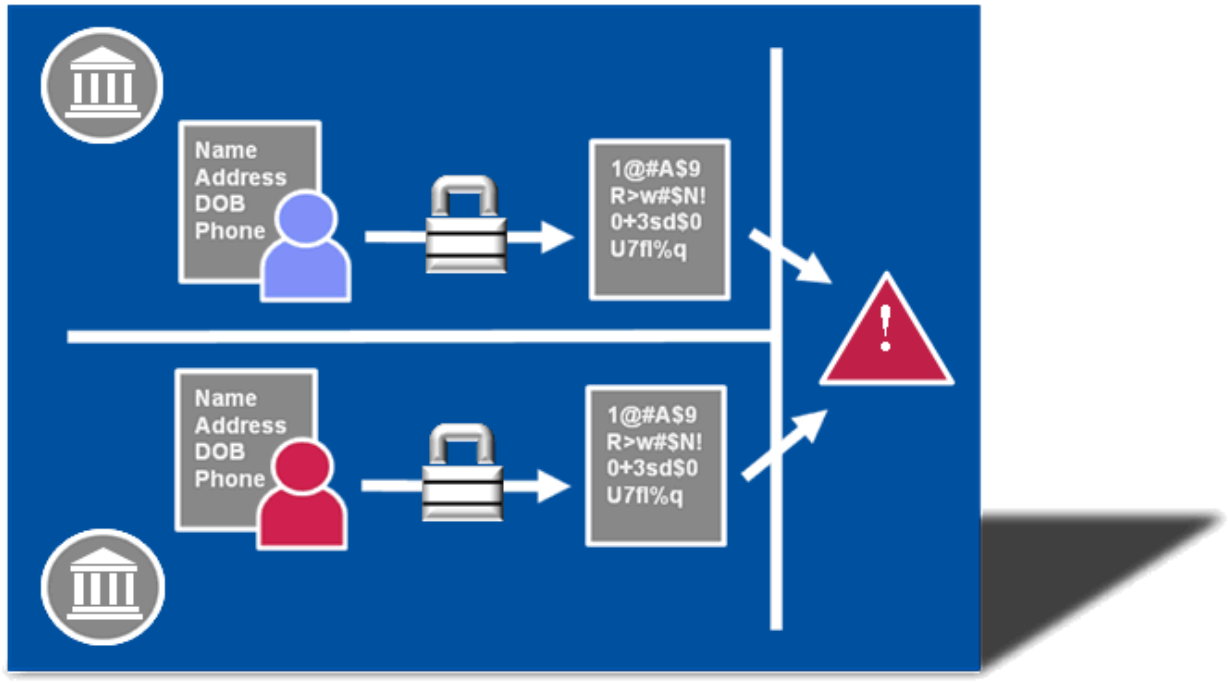
Source A	Source A (Encrypted)	Source B	Source B (Encrypted)	Match to Source B?
(Variant Created with IBM AR) Jonathan Smith	r1oBUKY?8t+E(a	(Variant Created with IBM AR) Jonathan Smith	r1oBUKY?8t+E(a	IBM AR: Yes

AR automatically links any encrypted variants back to the original source records. To reiterate, all the variant processing is done *at the originating site*, prior to the resolving process.

3. Anonymous Resolution Output

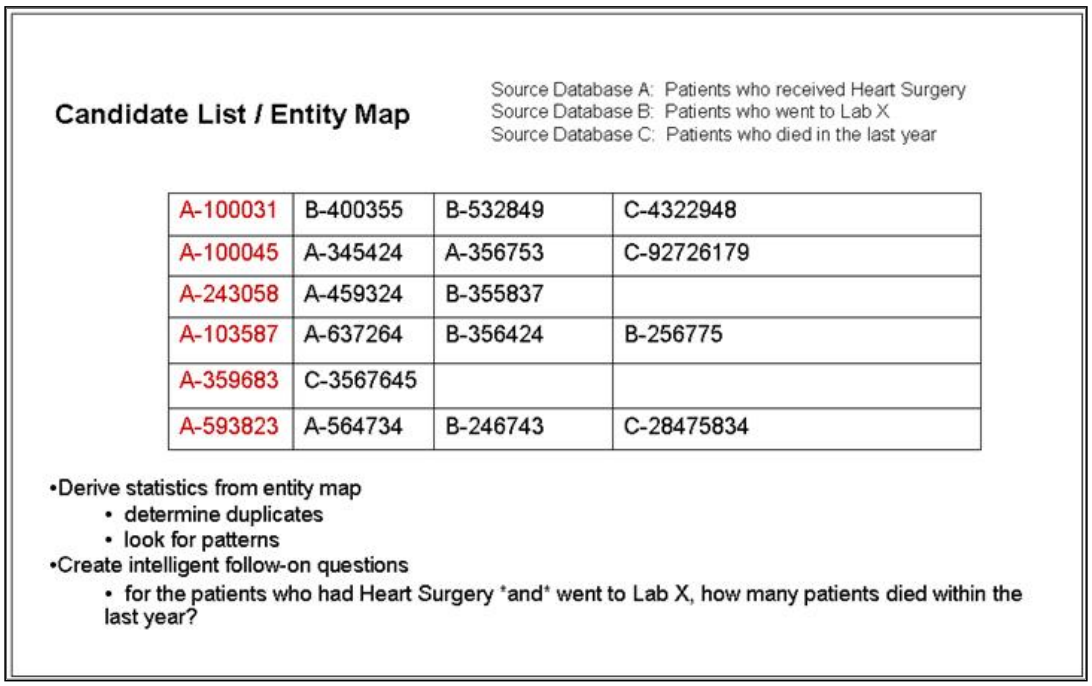
Once the source data has been standardized, enhanced, and encrypted, it is sent to an AR Resolver site. This site, typically in another physical location, serves to correlate the anonymized data and create the AR output. This process is shown in Figure 4.

Figure 4. Multiple Sites Contribute Anonymized Data to the Neutral Resolver Site



The AR Resolver correlates the encrypted records from multiple source systems. The AR Console is used to manage the matches. The output of the AR Resolver system is a correlated data entity mapping, shown in Figure 5. The entity map can be used to perform analysis on trends and participation across datasets. This data represents the relationships between the datasets.

Figure 5. Example AR Data Output



Consumers of the AR output are never able to see plain text forms of the data. The AR results are completely free of any personally identifiable information, as shown in Figure 5. Since the output information is being produced from an anonymized dataset, the only information available to consumers of the AR output is the list of match-pairs across the data sources.

III. ACCURACY ANALYSIS

A. OVERVIEW

The accuracy analysis compared matches obtained from the AR software to the gold standard matches obtained from the QS software. The hypothesis underlying this project is that, even though AR and QS take different approaches to matching records that originate from two separate data sources, they are expected to render similar results. We began by reformatting the data so that matches were represented by two columns of data: an ID for the HC dataset and an ID for the HSS dataset. An equivalent match was thus represented by a row in the QS dataset with HSS and HC IDs that each matched those IDs in the AR dataset.[13] A guide to the data format is displayed in Table 4.

[13] Note that all ID numbers were randomized before data processing began such that the IDs were not identifiable information.

Table 4. Output Format

Quality Stage	
Health Core Records	HSS Records
HC ID #	HSS ID#
HC ID #	HSS ID#

Anonymous Resolution	
Health Core Records	HSS Records
HC ID #	HSS ID#
HC ID#	HSS ID#

To facilitate processing, we concatenated the two IDs in each row to a single value. As such, each match then had a unique ID. A comparison of the single new match ID would let us ascertain whether a match appeared in one or both datasets. An example is shown in Table 5.

Table 5. Unique Match ID

Quality Stage	
Health Core Records	HSS Records
HC12345	HSS9876
HC33445	HSS9900

Anonymous Resolution	
Health Core Records	HSS Records
HC12345	HSS9876
HC112233	HSS8765

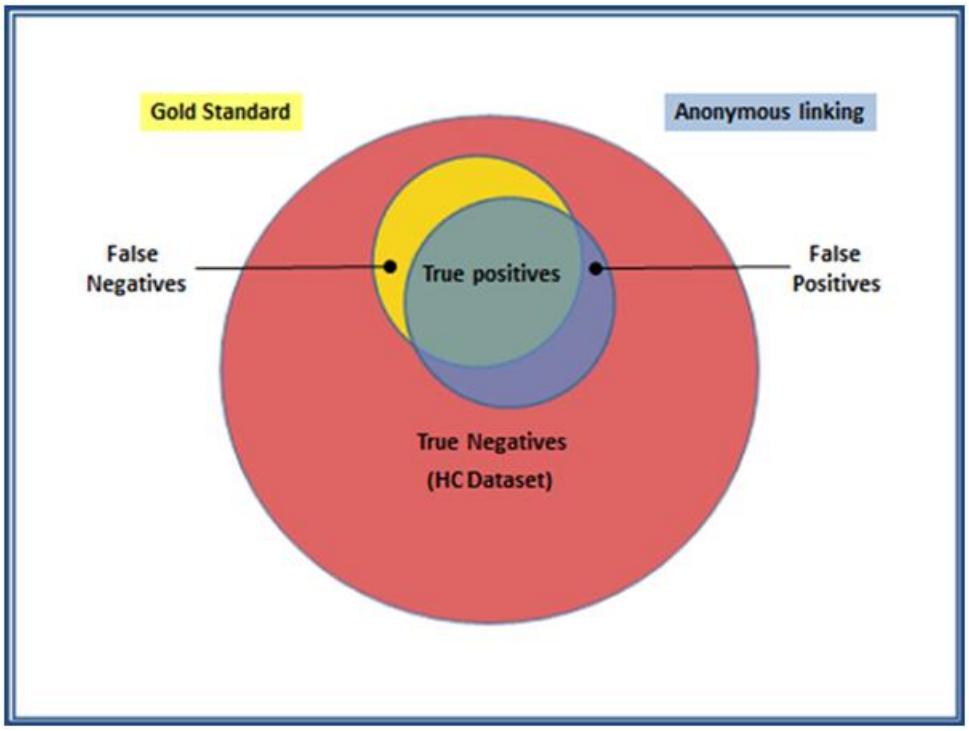
HC12345<->HSS9876
HC33445<->HSS9900

HC12345<->HSS9876
HC112233<->HSS8765

The following computations were done to evaluate the matching processes, as shown in Figure 6.

- The number of patients in the AR dataset who also appear in the QS dataset. These were expected to be true positive matches.
- The number of patients who appear in the AR dataset but who do not appear in the QS dataset. These were expected to be false positives.
- The number of patients who appear in the QS dataset but not in the AR dataset. These were expected to be false negatives.
- The number of patients from the HC dataset who did not appear in the QS dataset. Those who also did not appear in the AR dataset were expected to be the true negatives.

Figure 6. Accuracy Analysis Plan Diagram



B. STATISTICAL CONSIDERATIONS AND ANALYSIS

In order to quantify the statistical uncertainty of the accuracy parameters obtained by comparing AR matches to those from QS, we calculated the sensitivity, specificity, and positive predictive value. Where needed, we employed the widely-used R open source software version 2.15 and the "caret" (Classification and Regression Training) package.[14]

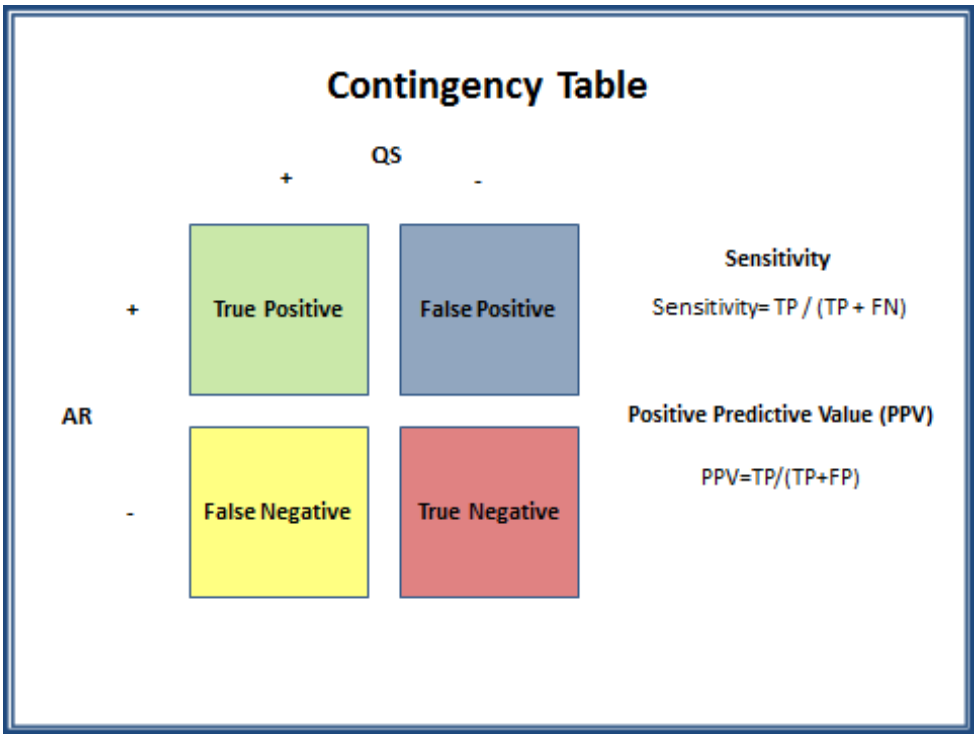
The sensitivity, or true positive rate, is a measure of how well the "test", in this case AR, correctly identifies those individuals who are in both datasets as determined by the gold standard, QS. It is calculated by dividing the number of true positives identified by AR by the total number of individuals who are present in both datasets as defined by QS (i.e., true positives and false negatives). The specificity, or true negative rate, is the accuracy of AR for determining that a match between the two datasets was not present, or one minus the false positive rate. The specificity is the number of false positives divided by the total number of individuals who are members of HealthCore but not also patients in the HSS registry (false positives and true negatives). *Note that due to the large size of the HealthCore dataset (approximately 43 million records) relative to the size of the HSS dataset (approximately 20 thousand records), specificities were universally high and are therefore not further reported.* The positive predictive value (PPV) is the likelihood or probability that a match identified by AR

[14] Both the R open source software and the Classification and Regression Training package are available at http://cran.r-project.org

was actually the same individual in both datasets, as determined by QS, or the number of true positives divided by the total of true and false positives identified by AR.

The positive and negative predictive values are derived from the resulting 2x2 contingency table that provides the numbers for Figure 6. These values depend not only on the sensitivity and specificity (calculated using the numbers of true and false results described above), but also on the number of individuals who actually are in the HSS registry and in the HealthCore dataset (analogous to disease prevalence in clinical prediction based on test results). A diagram of the contingency table is shown in Figure 7.

Figure 7. Contingency Table Diagram



IV. RESULTS

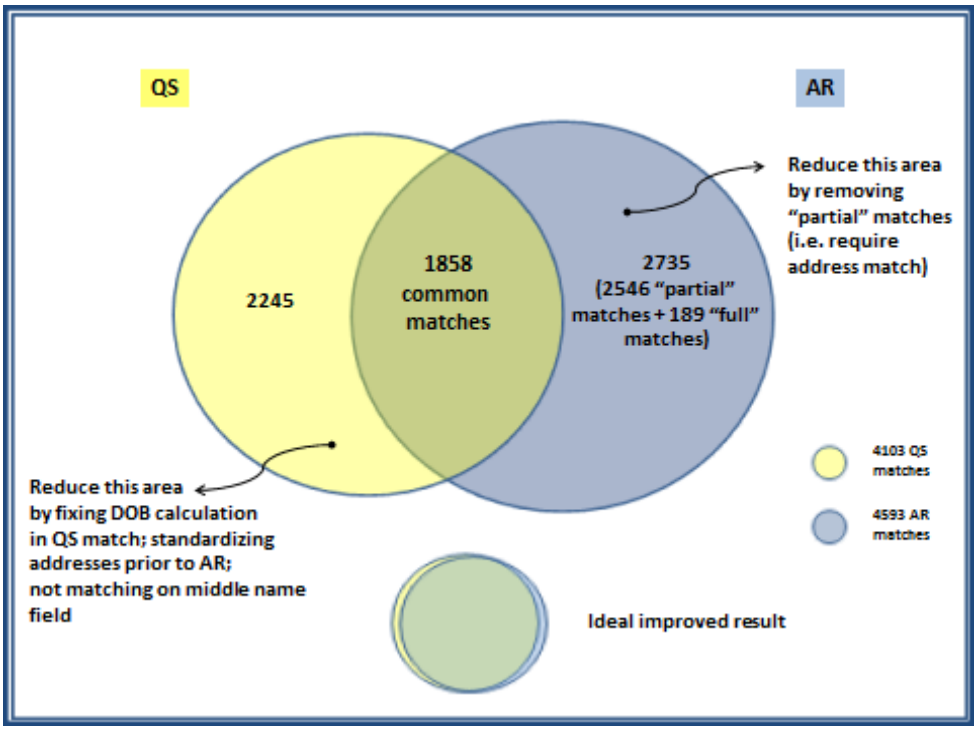
The hypothesis underlying this project is that, even though QS and AR take different approaches to matching records that originate from two separate data sources, they are expected to render similar results. As described below, the first attempt at matching the registry records from HSS with the claims data from HC resulted in incomplete matching in both QS and AR. However, there were readily addressable corrections to the procedures used that could be expected to yield improved results. Consequently, both QS and AR matches were repeated after implementing the improved preprocessing steps and correlation rules identified after the initial incomplete matches. The results of the first matching process and corresponding accuracy analysis are presented below as Phase 1, along with a detailed exploration of the reasons for many of the false results. The results of the second matching process and corresponding accuracy analysis are presented below as Phase 2, along with a detailed exploration of the reasons for any remaining false results.

A. PHASE 1

1. Phase 1 Results

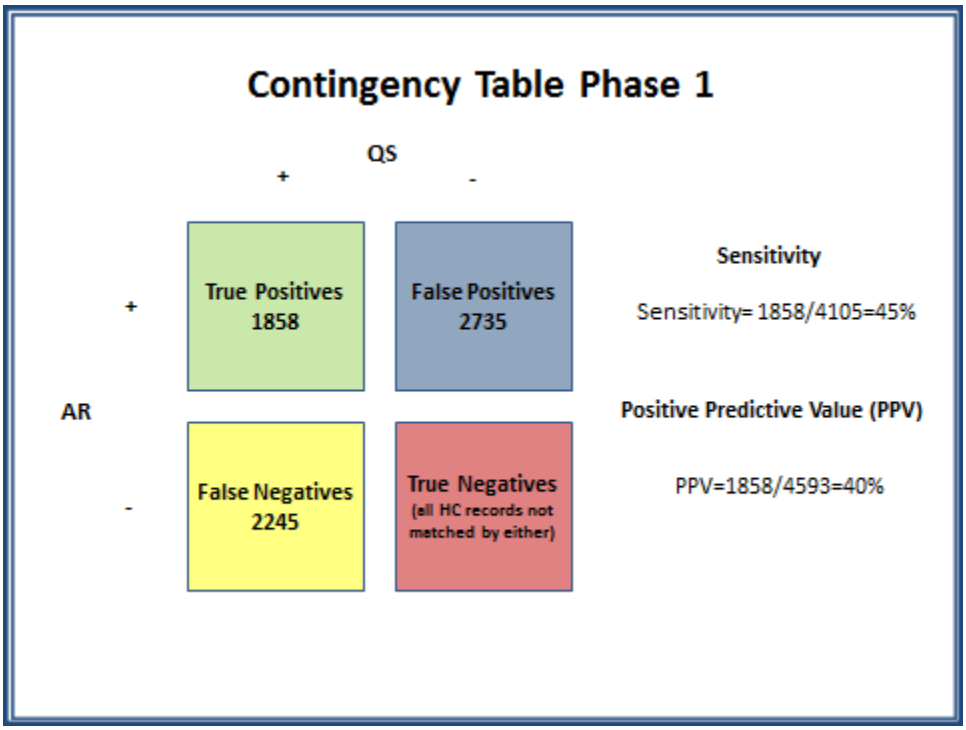
The Phase 1 QS and AR matching processes found 4,103 and 4,593 matches, respectively. Of these, 2,245 were only found in the QS match, while 2,735 were only found in the AR match. There were 1,858 matches in common. These results, along with the anticipated steps to improve the match, are shown in Figure 8.

Figure 8. Venn Diagram of Phase 1 Results



These findings are shown in the contingency table in Figure 9. The sensitivity was 45% and the PPV was 40%.

Figure 9. Contingency Table Analysis of Phase 1 QS and AR Matches



Based on the results of the initial accuracy analysis, the workgroup agreed that the results of the QS and AR matching processes were inadequate due to the large number of false negatives, and to a lesser degree, false positives, that resulted from AR in comparison to the matches observed in QS. It became clear that each program had used different preprocessing steps and different correlation rules.

For these reasons, it was determined that a second set of QS and AR matching processes should be run with more consistent preprocessing steps and correlation rules.

2. False Negatives and False Positives

The results indicated a surprisingly low proportion of common matches across the QS and AR processes. Consequently, in order to determine the reasons for this unexpected result, we initially examined a sample of ten matches in each of the false negative and false positive categories to identify patterns with regard to the frequency of false results. We then conducted more in-depth analysis to identify reasons for mismatches that could potentially be eliminated, leading to stronger agreement between the two matching procedures.

a. Initial False Negative Review

We examined a sample of ten false negative records that were found by QS but not AR. We identified the following itemized issues among these records:

1. Address standardization problem. Lane vs. “ln”.
2. AR did not make this match happen. We were unable to determine why the HC record was not found in AR.
3. Middle initial problem. HC record had a middle initial, where the HSS record did not.
4. True Negative. First name didn’t match in both records. Last name same. DOB totally different. QS not valuing DOB properly.
5. Initials in middle name problem.
6. Address standardization problem. Court vs. “ct”.
7. Address standardization problem. Street vs. “st”.
8. QS DOB problem. Should be a true negative.
9. Address standardization problem. E vs. East.
10. Address standardization problem. CT vs. Court.

b. In-depth False Negative Review

A majority of the time dedicated to explaining the low percentage of common matches was spent analyzing the QS-only matches (“AR False Negatives”), as this was the source of the largest discrepancy between the two result sets. We reviewed as many match-pairs as possible to get solid statistics reflecting the impact of each issue. We started with the set of 2245 QS-only matches. In examining these closely, we identified the following problems and solutions:

Bad Name Matches (10%): The QS configuration put too little weight on a solid name-match. This resulted in match-pairs that scored high in address, but had very low-score name matches. These should have been true negatives. For example: “Bob” matched “William”. This occurred both in the given name and surname fields. We examined all 2245 match pairs; 222 pairs fell into this category, representing about 10% of the QS-only matches.
Recommendation: QS should be configured to put a higher-value on name matches, even when the other match elements such as address are already strong.

DOB Issues (<1%): The QS-only output file did not carry with it the DOB information for the records. QS took into account the DOB during matching, but we were unable to examine DOB issues in the output because we did not have sufficient time to examine the substantial amount of original data.
Recommendation: QS should be reconfigured to place weight upon the DOB during the matching exercise. It should also be configured to output the DOB information in the match-results.

Middle Name Field (2%): AR did not match a record if one record had a middle name while the other did not. A setting within AR can be adjusted to eliminate this problem. All 2,245 QS-only matches were evaluated; 45 match-pairs fell into this category representing about 2% of the QS-only matches.
Recommendation: Middle name should be removed as a match element because it is highly unreliable in datasets.

Address Standardization/validation (88%): There was a clear pattern of missing address descriptors that caused AR to miss these matches. If one record had “123 main st” and the other record had “123 main”, AR did not consider this a match, while QS did consider it a match. Similarly, QS found “456 Broad parkway” and “456 Broad” as a match, where AR did not (perhaps it could have been “Broad Street”?). To assess how frequently this occurred, 500 records were examined. Of the 500 sample records, 441 records fell into this category, representing 88.2% of the QS-only matches.

Recommendation: Prior to anonymization, the entire address data, including zip code, should be 1) standardized (st=STREET) and 2) validated (an automated comparison against a real address to determine the validity of the address). This should resolve 88% of the discrepancies between the two matching sets.

Summation: During this close examination we accounted for nearly 100% of the discrepancies between the QS and AR matches with regard to "AR False Negatives"; there were a handful of other mismatch issues that comprised less than 1% of the total. As a result, the group determined that the AR system and pre-processing could be readily adjusted to enable significantly more valid matches, without also generating more false matches.

c. Initial False Positive Review

We examined a sample of ten false positive records that were found by AR but not QS. We identified the following itemized issues among these records:

1. AR partial match
2. True positive: QS not matching on multi-token last name
3. True positive: Missing zip code digit.
4. AR partial match
5. AR partial match
6. AR partial match
7. AR partial match
8. AR partial match
9. AR partial match
10. AR partial match

We determined that the AR partial matches were not aligned with the goal of this program. The AR partial matches did not require an address match and consequently were misaligned with the matching rules specified in QS. This accounted for 2,546, or 93%, of the 2,735 false-positive matches.

Recommendation: Because this definition of a match was not aligned with our project goals, the AR partial match category should not be counted as matches in Phase 2.

d. In-depth False Positive Review

AR-only matches ("AR False Positives"): Due to time constraints, we were not able to fully review the 189 match-pairs that were AR-only matches. Those reviewed appeared to be good matches, with strong name, address, and DOB matches. We could not explain why QS did not find these matches. The only aspect that looked common across these pairs was the presence of unusual names. The names seemed be very long and atypical, such as "Margarita Jenyansonashad" (a fictional example). Perhaps there is some logic embedded in the QS software which takes into account longer or less statistically common names. We did not have time to investigate this issue, therefore we are not sure whether or not to call these "False Positives". As they approximate 4.6% of the QS result set, we think it is reasonable to state that AR has 95%-100% precision.
Recommendation: When the results of Phase 2 are completed, a full analysis of the AR False Positive dataset should be done.

3. Summary of Findings and Recommendations

In this section we summarize our assessment of Phase 1 of the AR and QS results and provide recommendations for improving the results of Phase 2.

Matching rules: Although AR and QS take different approaches to scoring and matching records, they should produce approximately the same results. Both programs can be modified to use different weights pre-processing steps, and rules during their matching processes. Upon analysis, there were different pre-processing steps and matching rules used by QS than used by AR. Adjusting these rules should substantially improve the results.
Recommendation: Both products should be configured to run with similar rules for name, DOB, and address fields.

Middle names and middle initials: Middle names are notoriously unreliable data elements in name fields. It is impossible to determine whether a data source will reliably have middle name data available for matching. AR, because it uses a hashing method, treats middle names and middle initials as an important differentiator between records. QS puts little value on middle name data, essentially ignoring its presence or absence. This causes AR to under-match on these records.
Recommendation: Middle name data should be removed during the pre-processing step before anonymization. This should result in a large number of the QS-only matches moving into the "common" result set.

Match weights on full-name and DOB: QS was configured to place high weights on last name and address, and little weight on first name and DOB. In fact, there were cases in which DOB was ignored during the match, resulting in overmatching on many records.
Recommendation: QS should be re-run using DOB as a high-value differentiator. Also, first-name should have a higher matching weight. This should reduce the number of QS-only matches and thereby decrease the number of false positives for AR.

Address Standardization: QS uses address standardization to improve matching results. For example, it will reformat "main st" to "Main Street". Doing this in all datasets prior to running the matching process improves the chances of matching. Similarly, AR also has an address standardization capability, but this step was not run on the data prior to the test. This caused AR to fail to recognize matches like "123 circus ct." and "123 circus court".
Recommendation: Address standardization should be performed for AR prior to the anonymization/hashing step. This should result in more matches that are common between AR and QS thereby reducing the QS-only result set and the number of AR false positives.

Duplicates: Upon examination, the duplicates were contributed by the original source data, not the AR or QS process.
Recommendation: Duplicates should be resolved during the matching process.

B. PHASE 2

1. Terminology Clarifications

As noted, during the analysis of results in Phase 1 it became clear that certain terms were not well defined or used consistently. The group agreed with the following guidelines:

a. Quality Stage Output

High, Medium, and Low labels were arbitrarily set based on the total match weight, as follows. The values represent ranked assessments of the probability of the match.

- **High Probability:** weight >= 85%
- **Medium Probability:** weight between 75% and 84% inclusive (>=75 and < 85)
- **Low Probability:** weight between 65% and 74% inclusive (>=65 and <75)
- **Review:** weight < 65%

QS uses probabilistic matching algorithms, which derive statistics from the composition of the data itself. Therefore, common match values may rank lower than uncommon ones, since there is a higher probability of a random (incorrect) match among common values. For example, even though they are both exact matches, “John Smith” matching “John Smith” will have a lower match score than “Xavier Bouvier” matching “Xavier Bouvier”. This is because “John Smith” is a relatively common name and therefore does not weigh as heavily as a unique descriptor. In order to conduct the statistical analysis, all QS matches, regardless of their associated match scores, were included as the result of the gold standard matching process to which AR was compared. The labels of “High”, “Med”, and “Low” were deemed as not useful for determining levels of matching certainty.

AR uses deterministic matching algorithms, which give equal weight to equal variations regardless of the commonality of the data. For example “John Smith” matching “John Smith” will receive equal weight as “Xavier Bouvier” to “Xavier Bourvier” as they are both 100% matches. The AR deterministic approach is appropriate due to the fact it is matching hash values and cannot determine the statistical makeup of the original databases.

In order to conduct the statistical analysis, all QS matches, ”High” “Med” and “Low”, were included as the result of the gold standard matching process to which AR was compared. In essence, the labels of “High”, “Med”, and “Low” were considered not meaningful when compared to the AR matches.

The difference in matching approaches should be seen as a positive attribute in this test. The QS superior method of using statistical probabilities will create a superior level of matching. The AR results will be measured against this superior technique.

b. Anonymous Resolution Output

AR Full Match: In the Phase 1 output, AR matches contained a descriptor of “Exact Match”, even though the output includes variations that are very close but not exact matches. AR rules used to generate this set include:

- Exact Name, Address Hash and Exact DOB

- Exact Name, Address Hash and DOB(Swapped)
- Exact Name (First/Last Name Hash), Address Hash & Exact DOB
- Exact Name (First/Last Name Hash), Address Hash & DOB (Swapped)
- Close Name, Address Hash & Exact DOB
- Close Name, Address Hash & DOB(Swapped)

Note: Optimal rules above direct AR to create additional encrypted values for "Close" or "Swapped" values. For example, "Close Name" will create additional hash records for a name like "Jonny Smith" (Jonathan Smith). "Swapped DOB" will create additional hash records for the value (11/23/1965 could generate 11/32/1965).

We adopted the term "AR Full Match" to more accurately describe this set, in order to ensure the understanding that non-exact matches will be included.

AR Partial Match: In the Phase 1 output AR matches contained a descriptor of "AR Fuzzy Match" which suggested that there isn't an actual complete match. More accurately it included Name Key (First & Last Name concatenated without a white space) & DOB. It required no matching of an address field. This means it only matched on two of the three key criteria (ignoring address). To clarify the category, we decided to label this set of matches "AR partial matches" and did not include this set as an AR match in the Phase 2 results analysis, as it was not consistent with the specifications of the project to use all three. By definition, QS will NOT get these matches, as QS is looking for three items to match (name, DOB, and address).

2. Configuration Changes

The results of Phase 2 QS and AR matches were based on the following configuration changes:

a. Quality Stage

The QS software was reconfigured as follows:

- Standardization was performed for all input names to be treated as individuals. (QS is set by default to determine whether a name refers to an organization or an individual). Setting the configuration to individuals-only allowed for greater accuracy. This placed greater emphasis on weights for name and DOB.
- Dates were required, if populated, to be either identical or close. The previous data-run had not placed a weight on dates being identical, causing over-generation of matches.

b. Anonymous Resolution

The AR software was reconfigured as follows:

- Addresses were standardized by the same program used by QS before loading into AR. This ensured that both engines were examining the same address data.
- A rule was removed that had allowed for non-address match. In Phase 1, we allowed a category called "AR Partial Matches" containing strong Name+DOB without a requirement for an address match. This did not match the goals of the project and differed significantly from the QS configuration. In Phase 2, we configured the system to

expect a good address match in all matches, and to include exact matches and acceptable variations.

- Rules used included:
  - Exact Name Address Hash and DOB
  - Exact Name Hash Address and DOB (Swapped)
  - Exact Name Address and DOB
  - Close Name Address and DOB
  - Close Name Address and DOB (Swapped)

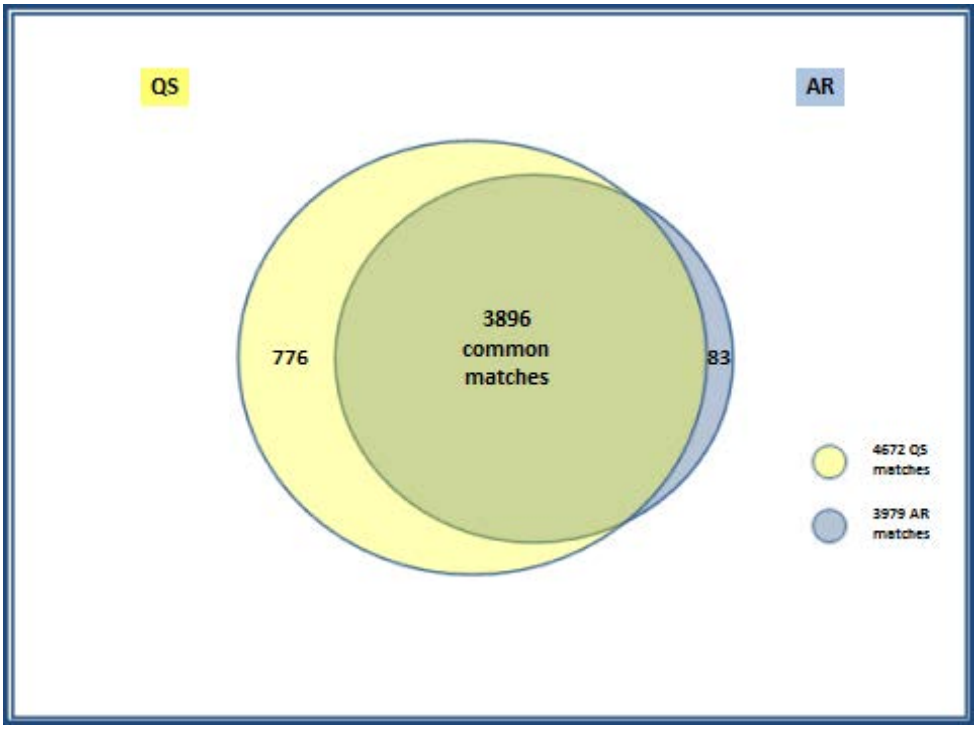
- AR Performance. We reconfigured the AR program to take optimum advantage of the hard drive available. This significantly reduced the loading time from 8 weeks to 2 weeks. Further performance improvements would be expected with additional CPUs.

3. Phase 2 Results

a. Overall Results

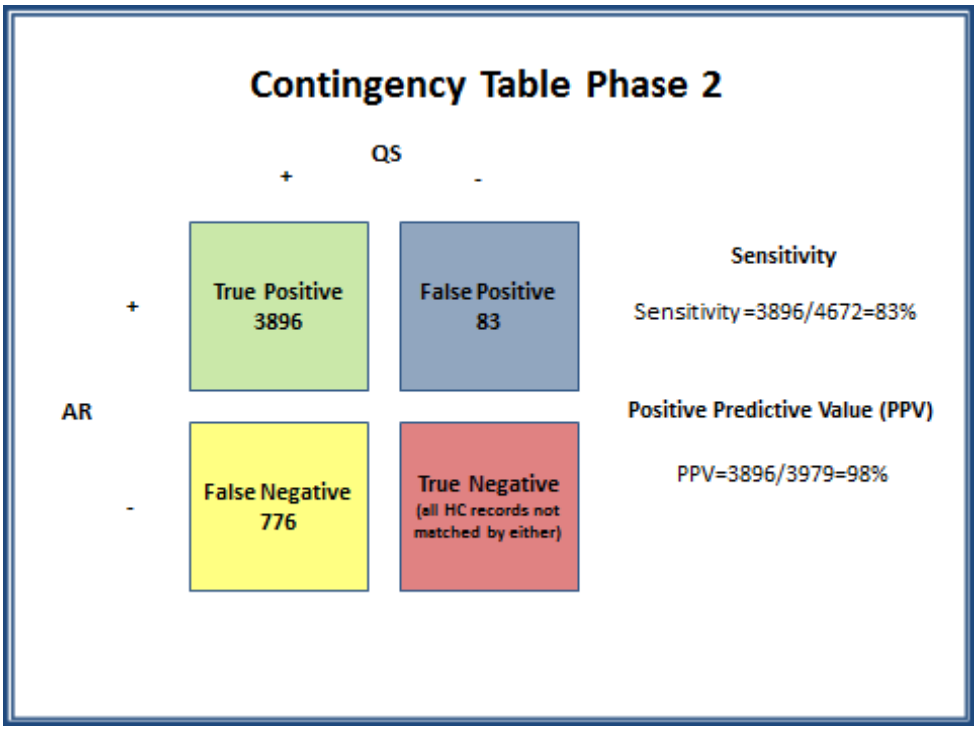
The second QS and AR processes found 4,672 and 3,979 matches, respectively. Of these, 776 were only found in the QS match, while 83 were only found in the AR match. There were 3,896 matches in common. These results are shown in Figure 10.

Figure 10. Venn Diagram of Phase 2 QS and AR Matches



These findings are shown in the contingency table in Figure 11. The sensitivity was 83% and the PPV was 98%.

Figure 11. Contingency Table Analysis of Phase 2 Results



b. Quality Stage Only Matches

As noted above, 776 matches were found by QS but not AR. Upon examination, AR failed to find these matches for the following reasons:

- 4 (0.5%) were found to be "true" false matches.
- 752 (97%) were missed by AR due to address validation issues. Address validation issues include misspelled street names, concatenated addresses, and other similar discrepancies. QS matched addresses with white space variations as well as extra information like apartment numbers. In contrast, the address matching rules for AR were more conservative. If the records were off by a misspelled street name or a missing the apartment number, AR did not count this as a match. Performing address validation would significantly improve this result. At the time of Phase 2, the team did not have address validation technology available to perform this function.

  - Example 1 (Whitespace issue = 10% of cases / 75 matches)

    HSS record has "Bushytail RD" as the street name
    HC record has "Bushy tail RD" as the street name
    QS matched this and AR did not

In this example, Address Standardization would change "RD" to "Road". However, it does not know that Bushtail should be all one word. An Address Validation process is required to confirm the correct spelling of the street name.

- **o** Example 2 (Address issues = 90% of cases / 677 matches)

  HSS record has "Unit 2" in address
  HD record does not have a Unit number in the address
  QS matched this and AR did not

  Both AR and QS can be configured to accept or ignore the Unit or Apartment number fields. Additionally, Address Validation could be used to supplement address information with unit or apartment numbers.

- 20 (2.6%) were missed by AR due to matching on names with initials. AR was not configured to match on initials. In general, because initials could generate a large number of false-positives in anonymous matching, we recommend using other attributes (such as SSN) to match if available.

c. Anonymous Resolution Only Matches

As noted above, 83 matches were found by AR but not QS. Upon examination, QS failed to find these matches for the following reasons:

- 42 (50%) were true matches, but were missed by QS due to AR's "fuzzy" DOB matching capability. AR was configured to allow for transposed or close date matches, while QS had a stricter DOB rule.

  - **o** Example 3

    HSS has DOB year of 1965
    HC has DOB year of 1966
    AR matched on this but QS did not

  - **o** Example 4

    HSS has DOB of 06/10/65
    AR has DOB of 06/01/65
    AR matched on this but QS did not

    In this example, AR is allowing a transposition of the day-field within the date. This is accomplished by generating an additional encrypted key at pre-processing time. This additional key can be matched during resolution and mapped back to the original record ID.
    Recommendation: A decision about what level of date variation is optimal can be accommodated by both systems. It is common to allow some switches among day, month, and year due to the possibility of data entry variations.

- 8 (10%) were address standardization/validation issues. Many of these were true matches.

  o Example 5

  HSS had “123 Main LA”
  HC had “123 Main Lane”
  AR matched this but QS did not

  In this example, the address standardization did not reformat LA to Lane. LA is an ambiguous abbreviation that could have meant “Lake”. Therefore, the QS engine missed the match.
  Recommendation: Using address validation in the preprocessing step would correct this issue.

- 33 (40%) were word similarities that QS did not find. Many of these were true matches, but were difficult to quantify as the hash codes were determined by the type of error in the data.

  o Example 6

  HSS had “Randol Street”
  HC had “Randolf Street”

  AR matched on this but QS did not.
  Recommendation: Address validation would have helped in this case.

4. Summary of Findings and Recommendations

Address Standardization:, Address standardization is the process of converting addresses to a standard representation, for example, converting “123 Main st.” to “123 Main Street”. Based upon findings in Phase 1 and Phase 2, address standardization improves matching by as much as 35%.
Recommendation: All contributors should standardize their addresses with the same standardization technology prior to anonymizing the records.

Address Validation: Address validation is the process of ensuring an address actually exists. It can also be used to enhance the address record with additional information. For example, “123 Bushy tail Street, leesville, tx” could become “123 Bushytail street, Leesville, TX, 99999”. Address validation can also identify erroneous records that should be examined prior to anonymization. It appears that address validation could solve as much as 97% of the false negatives (found by QS only) found in Phase 2. It also could correct up to 10% of the false positives (found by AR only) found in Phase 2.
Recommendation: All contributors should validate their addresses with the same validation technology prior to anonymizing the records.

Decisions on matching criteria: It is important to decide on matching criteria appropriate to the task at hand. The goals of certain programs might be suited to over-matching (i.e., false negatives are a problem). Other programs might be suited to conservative results (i.e., false positives are a problem).

The resources available to review and evaluate true and false matches to adjust the matching criteria for optimal results should also be considered.
Recommendation: Matching criteria appropriate to the task at hand should be decided upon in advance. If expanded results are desired, then the level of "fuzzy" matching can be increased to find all possible matches. If there is a concern regarding false identification, then a more conservative configuration can be used for the matching. That is, priority can be given to sensitivity at the expense of specificity (or vice versa) according to the needs of the tasks for which the matching is implemented.

Additional matching attributes will improve results: In this test, we used only three matching attributes: name, date of birth, and address. The addition of even one more identifying field, such insurance ID, would contribute significantly to the accuracy of the matches. The inclusion of additional fields effectively lessens the considerable degree of weight placed on the address field, which often contains unpredictable variations. If additional data is present in a record, it can be used to further enhance matching accuracy. While social security numbers are uniquely useful matching attributes, they are generally unavailable in claims data.
Recommendation: Additional matching attributes should be used if they are populated consistently and accurately.

V. CONCLUSIONS AND IMPLICATIONS

Anonymous linkage of individual records from a medical device registry to a claims database can be done with relative feasibility using IBM's AR software.

In Phase 2 of this study, the matched records obtained by anonymously linking databases using AR were compared to the matches obtained using fully identified data. This "gold standard" was achieved through the use of another product, IBM's QS software. The statistical analysis of this comparison demonstrated a high level of accuracy obtained using AR. The results indicated that 3,896 of 4,672 records identified as matches by QS were also matched anonymously by AR, for a matching sensitivity of 83%. Only 83 of the 3,979 matches by AR were incorrect (i.e., two different individuals rather than the same person) leading to a positive predictive value for AR of 98% (3,896/3,979=98%). Further, upon closer examination, 42 of the 83 AR matches labeled as incorrect appeared to be true matches. These were missed by QS due to its stricter matching rules but caught by AR due to its "fuzzy" matching capability.

These findings indicate that almost all true and valid matches existing in both databases can be found and linked anonymously with very few people who are not true matches identified as such by the anonymous linkage software. It is important to note that AR software settings can be adjusted by establishing looser or tighter matching criteria which will result in higher or lower sensitivity at the expense of specificity and, thus, of the positive predictive value.

Limitations of this study include:

1. The results reported required two rounds of matching in order to discover sources of errors during an initial attempt to match the datasets. Some were related to inconsistencies between how the identifying information was coded in the databases. Others were because of inconsistencies between the two software packages in the way the matching criteria were set. These issues, once identified, were readily corrected by

adjusting the software settings or via the use of additional software to standardize and validate addresses and to standardize how the date of birth was recorded.

2. What was learned about data standardization and how to correct these inconsistencies could be incorporated into the matching procedures going forward.

3. The speed at which the matching process can be conducted relies heavily on the storage capacity of the hardware in which the databases are housed. We were able to accomplish a higher speed of matching during our second round (Phase 2) by improving the availability of storage compared to the first round (Phase 1). This should be taken into consideration in future tests and analyses.

Implications of this study include:

1. This initial exercise suggests that a full scale implementation of anonymous linkage should be considered in order to test, under actual operational conditions, whether such a procedure remains feasible, and whether it produces valid and usable results. This would require configuring AR based on the recommendations above and applying it to match individuals from these two separate data sources including additional information about health care services and outcomes. Feasibility could be assessed via quantification of the amount of time and resources required to do the matching in this operational environment. Validity and usability could be evaluated without necessitating the application of another fully identified gold standard match, as was done in this experiment. By implementing previously identified best practices, the AR match could be done in a single phase.

2. If the above exercise proves successful, then there would be strong evidence that individuals from two data sources can be linked for the purposes of medical product safety surveillance in a feasible and accurate manner without sharing PHI.

3. Linking individuals from two complementary data sources could greatly enhance the scope and completeness of the data available for medical product effectiveness and safely surveillance, and hence, the feasibility and value of post-marketing surveillance of drugs and medical devices.

VI. LITERATURE REVIEW BIBLIOGRAPHY

1. Thelot B, Chevallier J. [Numeration of patients from anonymous data: a method established with the epidemiological data of public health of Paris]. Revue d'epidemiologie et de sante publique 1988;36:226-34.
2. Thelot B. [A general solution to the linkage of anonymous medical data]. Comptes rendus de l'Academie des sciences Serie III, Sciences de la vie 1990;310:333-8.
3. Keller JE, Howe HL, Noak JR. An algorithm for matching anonymous hospital discharge records used in occupational-disease surveillance - anonymous record matching algorithm. Am J Ind Med 1991;20:657-61.
4. Jaro MA. Probabilistic linkage of large public-health data files. Statistics in medicine 1995;14:491-8.
5. Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. Statistics in medicine 1995;14:499-509.
6. Bouzelat H, Quantin C, Dusserre L. Extraction and anonymity protocol of medical file. Proceedings : a conference of the American Medical Informatics Association / AMIA Annual Fall Symposium AMIA Fall Symposium 1996:323-7.
7. Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up : quality assessment of an anonymous record linkage procedure. Int J Med Inform 1998;49:117-22.
8. Quantin C, Kerkri E, Allaert FA, Bouzelat H, Dusserre L. Security aspects of medical file regrouping for the epidemiological follow-up. Studies in health technology and informatics 1998;52 Pt 2:1135-7.
9. Ades AE, Walker J, Botting B, Parker S, Cubitt D, Jones R. Effect of the worldwide epidemic on HIV prevalence in the United Kingdom: record linkage in anonymous neonatal seroprevalence surveys. Aids 1999;13:2437-43.
10. Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D, Humaine TCEGCdIedSdII. Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990–1993. International Journal of Epidemiology 2000;29:168-74.
11. Blakely T, Woodward A, Salmond C. Anonymous linkage of New Zealand mortality and Census data. Australian and New Zealand journal of public health 2000;24:92-5.
12. Borst F, Allaert FA, Quantin C. The Swiss solution for anonymously chaining patient files. Studies in health technology and informatics 2001;84:1239-41.
13. Bradley CJ, Given CW, Luo ZH, Roberts C, Copeland G, Virnig BA. Medicaid, Medicare, and the Michigan Tumor Registry: A linkage strategy. Med Decis Mak 2007;27:352-63.
14. Kijsanayotin B, Speedie SM, Connelly DP. Linking patients' records across organizations while maintaining anonymity. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2007:1008.
15. Trombert-Paviot B, Couris CM, Couray-Targe S, Rodrigues JM, Colin C, Schott AM. Quality and usefulness of an anonymous unique personal identifier to link hospital stays recorded in French claims databases. Rev Epidemiol Sante Publique 2007;55:203-11.
16. Weerasinghe D, Rajarajan M, Elmufti K, Rakocevic V. Patient privacy protection using anonymous access control techniques. Methods of information in medicine 2008;47:235-40.

17. Quantin C, Fassa M, Coatrieux G, Riandey B, Trouessin G, Allaert FA. Linking anonymous databases for national and international multicenter epidemiological studies: A cryptographic algorithm. Rev Epidemiol Sante Publique 2009;57:33-9.
18. Kimura S, Sato T, Ikeda S, Noda M, Nakayama T. Development of a database of health insurance claims: standardization of disease classifications and anonymous record linkage. Journal of epidemiology / Japan Epidemiological Association 2010;20:413-9.
19. de Lusignan S, Navarro R, Chan T, Parry G, Dent-Brown K, Kendrick T. Detecting referral and selection bias by the anonymous linkage of practice, hospital and clinic data using Secure and Private Record Linkage (SAPREL): case study from the evaluation of the Improved Access to Psychological Therapy (IAPT) service. BMC Med Inform Decis Mak 2011;11.