IMMUNIZATION INFORMATION SYSTEMS
PATIENT-LEVEL DE-DUPLICATION BEST PRACTICES

National Center for Immunization and Respiratory Disease (NCIRD)
Informatics and Data Analytics Branch (IDAB)

February 6, 2025

1 Executive Summary

Patient-level de-duplication (also called patient matching, patient de-duplication, or patient identity management) is the process of finding and removing redundant patient records from a database. Patient matching and patient de-duplication are essential data processing capabilities for Immunization Information Systems (IIS). These capabilities ensure that updates and queries apply only to the correct patient record and prevent fragmented and duplicate information from being added to an individual's health records. The inability to consistently determine which records represent the same patient and errors in combining the data contained in a patient's record negatively affect the overall data quality, usefulness, and credibility of public health immunization record keeping.

This document, a result of a CDC-sponsored project, is designed to be read by programmatic, technical, and operational experts who are involved in creating or maintaining an IIS. The document intends to bridge the gap between technical and program staff so they can have a mutual understanding of the issue of patient-level de-duplication and target actions to address these recommendations.

Best practice guidelines on patient-level de-duplication, documented within this report, will positively affect immunization registries by encouraging common de-duplication practices. This will thereby improve overall data quality and usefulness of registry information. The best practices guidelines are also technology-neutral and foster collaboration and communication amongst IIS professionals.

1.1 Findings and Best Practices

Detailed best practice guidance for day-to-day IIS patient de-duplication operations is outlined in both the foundational and advanced practice sections of this document. While specific best practice guidance has been summarized in the various practice-related sections of this report, it is believed that certain best practice information can broadly benefit the IIS national community and also help establish the discussion surrounding a long-term agenda.

Patient de-duplication is a multi-step process. Accordingly, best practices need to be understood within the context of a generalized process model. This model is presented in the body of the report. Currently, there is not a single idealized best practice process for patient-level de-duplication. There are wide variations in needs, capabilities, resources, and business practices; however, the expert panel believes that there are a number of techniques that can support patient-level de-duplication efficiencies in virtually all circumstances. While idealized process discussions were considered non-productive, certain techniques were identified that can enable productivity.

The following table summarizes the project's key findings and best practices by domain. Additional detailed information on the best practices can be found in section 5 of this document.

Domain	Findings	Best Practices
General	• The national practice community needs to begin anticipating how to best leverage their de-duplication processes into the evolving state and national health information technology architectures and in conjunction with Meaningful Use • Single, discontinuous efforts are not adequate to provide the functionality required to sustain continued improvements • Consequences of inappropriately merging the records of two patients are more severe than duplicating a patient's instance in the database	• Formally document all facets of patient de-duplication processes, including the business rules for each step of the matching and de-duplication process • Apply a business process approach to planning, implementing and documenting patient de-duplication practices • Understand the functional differences among de-duplication approaches for real-time, incoming, and retrospective processing and the strengths and weaknesses of each • Establish formal programs to routinely and systematically recognize and reward individuals and organizations who greatly assist IIS data quality at the jurisdictional level • Err on the side of preventing false patient record data merges and failing to match two records for the same patient (also called false negatives) • Participate in the on-going patient de-duplication process improvement dialogue whether a technical or non-technical SME • Formalize a body of knowledge which can help further solidify implementations and drive efficiencies acceptable to the national IIS community • Move towards the establishment of a universal or national patient identifier • Implement better mechanisms for sharing and collaboration among IIS around de-duplication best practices • Continue additional sustained expert panel work
Software Approaches, Capabilities, Specifications	• Emerging role of Master Patient Indexes (MPIs) is uncertain relative to IIS patient de-duplication	• Have a greater understanding of de-duplication "black-box" operations and the deterministic and probabilistic techniques being used

Domain	Findings	Best Practices
and Measurement Metrics	• There is currently no standard road map for IIS integration into Health Information Exchanges (HIEs) or other arrangements integrating cross-jurisdictional health information • Purely deterministic implementations eventually hit a ceiling of diminishing marginal returns	• Participate in de-duplication engine set-up and ongoing reviews • Utilize active discussion and on-going review with SMEs and technical support to identify threshold scores in conjunction with the needs of local stakeholders, local constraints and available data • When evaluating de-duplication engines, look for the following functionality: o Recognize when records have previously been adjudicated, i.e., de-duplication software should prevent multiple redundant record reviews o Perform comprehensive edit checks on manual data entry to standardize data contained in the database o Evaluate incoming data for completeness, timeliness, and accuracy through online prompting and edit checks, pop-up windows and other automated techniques o Provide on-line help as well as suggestions regarding formatting during manual data entry process o Illustrate potential duplicate records during manual data entry o Merge and unmerge patient records in more standardized ways o Utilize well-developed blocking techniques with a high number of unique values to reduce the overall number of candidate pairs to evaluate o Utilize machine learning to allow for further sophistication in correctly identifying and maintaining patient records o Implement more probabilistic methods for increased volumes and more complex problems

Domain	Findings	Best Practices
		○ Adjust and configure algorithmic techniques and thresholds as data sets change and evolve ○ Use a combination of deterministic and probabilistic algorithmic methods ○ Apply advanced algorithms to process last name data (**see page 36)** ○ Utilize specific five measures of sensitivity, specificity, accuracy, precision, and false positive rate **see page 46)** to benefit practice efficiency ○ Utilize additional useful measures for understanding IIS operations and improving data quality (**see page 47 and 48)**
Incoming Data and Manual Data Entry	• Despite increasing automation, manual data entry remains an important method of data origination • Manual data entry and multiple data sources introduce variations in data, typographical errors, and data omissions, affecting overall data quality • Multi-tier approaches to public and private immunization provider data problems are needed • The order in which records are examined may influence the outcome of record comparisons in certain situations • A road map is needed to further the jurisdictional mapping of IIS data to NVAC cord data elements and functional standards	• Provide the broad community of immunization data providers, including HMOs, pediatric associations, schools, pharmacies, insurance companies, and other institutions with formal feedback regarding the data quality needs of IIS • Utilize fact sheets, FAQs, dedicated expert calls, user group exchange webinars, and web-based training to improve the quality of data from immunization data providers • Follow up on trends in data originating from provider interfaces and encourage providers to review and act upon response files and error messaging • Encourage providers to utilize standardized HL7 messages • Encourage providers to run Vaccine for Children (VFC) and Assessment reports • Do not accept data originating from a source that is not approved • Do not utilize information from schools or insurance companies • Provide well-documented options to providers for submission of their data

Domain	Findings	Best Practices
		• Train data entry users on the best search methods supported by the IIS and provide detailed documentation and training • Perform better screening and cleansing of incoming data, particularly regarding placeholder and missing data, to ensure that incoming data meet minimal processing requirements • Perform systematic testing of format and content of incoming data within the on-boarding process for a new data source • Implement the option of utilizing data fields such as birth order, race, and ethnicity for de-duplication processing • Use specific unique identifiers such as social security number (SSN), medical record number, chart number, or birth certificate number to quickly find a match in the IIS and prevent calling the de-duplication engine • Make business decisions not to utilize certain types of records that present themselves (e.g., baby boy or girl) • Identify life status changes during the manual data entry process to aid in the identification of situations creating duplicate records or fragmented histories • Apply standardization rules for each component of a client's name **(see page 35)** • Standardize address and phone number information through a detailed examination of available address components (**see page 36 and 37)**
Retrospective Processes	• Retrospective examination of IIS data to find duplicate patient records and other forms of data quality problems should be considered a universal best practice	• Perform periodic retrospective examination and de-duplication of IIS patient records*(critical)* • Actively monitor the results of retrospective processing as an important source for improvement

Domain	Findings	Best Practices
	• The processes used in retrospective patient de-duplication may be different than front-end de-duplication processes. IIS SMEs need to understand and actively manage these differences to improve data quality.	• Utilize automated approaches to review audit trail artifacts to provide useful metadata
Manual Review Processes	• Manual data review processes are expensive and time-consuming	• Utilize audit trails and manual review files to identify improvement opportunities • Utilize objective data quality measures • Perform systematic reviews of pending logs to identify recurrent problems and logic gaps • Research and access additional information to increase the likelihood of making accurate determinations and document these situations to provide insights into operational weaknesses • Agree upon SME and technical activities that can reduce the burden of manual review processes • Note and pass along new variations to technical personnel to incorporate into a new standardization process (e.g., MLK for Martin Luther King) • Identify and apply culture-specific conventions (e.g., family members sharing the same date of birth)
Testing	The expert panel notes that there has been significant evolution in IIS operations over the past decade and their findings and recommendations, with regard to the development of new IIS patient de-duplication test cases, can be found in a separate testing document (Volume 2).	

Table 1.1: Summary of Project's Key Findings

2 Background and Project Overview

Immunization Information Systems (IIS) are confidential, population-based, computerized information systems that collect vaccination data within a defined geographic area. IIS are an important tool to increase and sustain high vaccination coverage by consolidating vaccination records from multiple providers into a single immunization record.

The ability for physicians, hospitals, and other healthcare providers to send immunization records to IIS electronically is a key element of what has been termed “Meaningful Use.” Meaningful Use is the ability to exchange complete and accurate electronic patient information, based upon the set of standards defined by the Centers for Medicare & Medicaid Services (CMS), in a way that can improve healthcare efficiency and patient outcomes.

Meaningful Use is defined by using certified Electronic Heath Record (EHR) technology in a meaningful manner (for example electronic prescribing); ensuring that the certified EHR technology is connected in a way that provides for the electronic exchange of health information to improve the quality of care; and submitting information on quality of care and other measures to the Secretary of Health & Human Services (HHS). The sending of provider immunization data to public health jurisdictional IIS has been incentivized by federal legislation, namely the American Reinvestment & Recovery Act (ARRA) and Health Information Technology for Economic and Clinical Health Act (HITECH). Accordingly, the volume of patient records being sent electronically to jurisdictional IIS has increased dramatically and will continue to increase. Therefore, IIS are under pressure to improve their overall data quality programs.

The last formal examination of IIS patient-level de-duplication methods along with the development of patient de-duplication tools was performed by the Centers for Disease Control and Prevention (CDC) in 2002, and much has changed since that time. The CDC notes that, given the importance of IIS in national Meaningful Use objectives, there is much to be gained from a fresh examination of patient de-duplication best practices.

A new CDC-sponsored patient-level de-duplication project aspired to bring de-duplication processes up to a best practice standard by specifying the development of best practices accompanied by test cases to test for both sensitivity and specificity and other accuracy measures. Additionally, the project examined and proposed practice-based solutions for the use of IIS data in a Master Patient Index (MPI) or similar environment to allow de-duplication engines to yield better and more accurate results through the use of a clean and complete data set.

In summary, the project sought to accomplish the following:

- Streamline, standardize, and improve overall IIS patient de-duplication processes
- Increase IIS expertise in patient de-duplication best practices
- Improve patient de-duplication and data quality best practices, which can lead to improved single patient hit rates from patient query/response use cases
- Make recommendations on how to implement improvements to an IIS
- Create a standardized set of test cases that can be used across all IIS

2.1 Project Approach

To address the problem of duplicate patient records in IIS, the project established a Patient Data De-duplication Expert Panel. The panel consisted of 14 Subject Matter Experts (SMEs) and Expert Reviewers from the following organizations:

- American Immunization Registry Association (AIRA)
- Indian Health Service (IHS)
- EHR vendors
- IIS programs and vendors
- IIS consultants and de-duplication experts
- Academic institutions

The work of the expert panel was performed during the period of August, 2011 through March, 2013. Work was assigned to one of two roles: 1) SMEs for primary content generation and 2) Expert Reviewers for content and product review.

As an expert panel, the group represented decades of profound expertise in IIS patient de-duplication procedures, tools, methods, and system administration. The membership of the expert panel is detailed in Appendix A.

The expert panel agreed upon the following mission statement: "The Patient De-duplication Expert Panel is comprised of key stakeholders focused on developing best practices and resources to improve patient de-duplication processes and the quality of IIS patient data."

The expert panel work followed a well-defined approach:

- Panel Recruitment
- Off-Line Research and Preparation
- Panel Work
  - Phase 1: Orientation
  - Phase 2: Literature Review
  - Phase 3: Patient-Level De-Duplication National Practice Assessment (NPA) results and manuscript
  - Phase 4: Vocabulary Definition
  - Phase 5: New Test Case Specification Development
  - Phase 6: Best Practice Statement Development
  - Phase 7: Final Report Development and Dissemination
  - Phase 8: Test Case Development and Dissemination
- Conference Presentations
- Publications and Website Updates
- Lessons Learned

Orientation involved initial preparatory off-line work including literature reviews, assembly of pertinent materials, production of preparatory notes, analysis of processes, and development of preliminary drafts. This effort was performed by a small group of business analysts and SMEs.

The Patient-Level De-duplication NPA examined and reported trends, problems, and approaches currently being taken on a national basis. A peer-reviewed paper is the by-product of this effort.

The work of the expert panel was conducted utilizing formal project management and facilitation techniques. Work methods included facilitated, pre-scheduled, bi-weekly teleconferences with expert panel SMEs that involved the following:

- Sharing of individual experiences
- Group discussions of patient de-duplication issues
- Voting via SurveyMonkey to stimulate and elicit best-practice agreement and disagreement
- Development and review of numerous materials
- Drafting of consensus-based recommendations

The CDC sponsored an intensive, 3½-day in-person session in Atlanta, GA from February 21-24, 2012. This in-person meeting covered all of the domain areas in the scope of work, included the full workgroup of expert panel SMEs, and utilized facilitated modeling techniques. The development and formulation of consensus-based recommendations occurred during strategic group breakout sessions.

The post in-person session work finalized the development of the best practice discussions and test case specifications. Additional teleconferences were dedicated to reviews of specific patient de-duplication practice questions by dividing up the work for development by small groups of SMEs and then by the group in its entirety. The expert panel's definition of consensus did not reflect 100% agreement, but rather "I can live with that and support it."

To help organize and coordinate the expert panel's work, Northrop Grumman Corporation's (NGC) Public Health Division was retained as the project contractor. Northrop Grumman provided project management, IIS de-duplication subject matter expertise and administrative support. Their scope of work included recruiting and constituting the expert panel; providing guidance toward collaborative examination, evaluation, and analysis; facilitation services; and proposing practice-based standardized solutions. Additionally, NGC supported test case development and final report authorship and production.

2.2 CDC IIS Panel Scope of Work

The major focus of this project was the issue of IIS patient-level de-duplication. This focus included the development of best practice guidelines and the creation of an updated set of test cases.

De-duplication of immunization records can be a two-fold problem that includes de-duplication at the patient level (e.g. two records describe the same patient) and de-duplication at the vaccination event level (e.g. two records describe the same immunization event). The scope of this project considered only the first of these two processes. Additionally, the project was not focused on the assembly of lifetime immunization records or on clinical decisions related to the immunization schedule.

The expert panel focused on five domain areas:

1. **De-duplication software approaches, capabilities, specifications and measurement metrics**
    - Practice-based evaluation of the efficacy of de-duplication approaches
    - Validation of contextual models
    - Best practice guidance on the ability of de-duplication software to yield better and more accurate results

2. **Incoming data and manual data entry de-duplication practices**
    - Practice recommendations around the validation and cleansing of incoming data using unique identifiers to shortcut de-duplication process interrogation
    - Identification of the most problematic data sources and situations
    - Guidance on prescreening incoming records to reduce manual effort
    - Recommendations to external providers to procedurally avoid duplication situations

3. **Retrospective de-duplication processes**
    - Best practices around de-duplication of existing patient data
    - Specifications around what additional data elements, particularly from the immunization history, may be useful
    - Identification of idealized record merge and unmerge practices

4. **Manual de-duplication review processes**
    - Identification of strengths and weaknesses of approaches, processes, and techniques used
    - Consideration of merge and unmerge processes
    - Identification of manual review productivity improvements

5. **De-duplication testing**
    - Development of an updated and expanded set of test cases to help assess the ability of an IIS to detect and de-duplicate patient records
    - Specifications for more robust patient de-duplication test cases, including the considerations for measuring sensitivity and specificity

- Specifications for the nature, type, and volume of test cases
- Review of new test cases featuring updated and expanded data utilization
- Recommendations for test case packaging, distribution, and use

The scope of this document includes domains 1 through 4. The outputs of the Domain 5, De-Duplication Testing, are documented in a separate report (Test Case Development & Utilization) that will accompany the new test cases.

The expert panel developed a number of very detailed artifacts. Relevant materials have been collected for inclusion into this final report and its appendices. Other materials are referenced within this report and documented separately. These artifacts include:

- Comprehensive literature review
- Common IIS patient de-duplication vocabulary
- Patient-level De-duplication National Practice Assessment (NPA) results and manuscript
- Best practice consensus on incoming, retrospective, and manual de-duplication methods, including the use of specific de-duplication algorithms and techniques
- Review and analysis of the 2002 Patient De-duplication Test Toolkit
- New, more comprehensive, evidence-based series of patient de-duplication test cases that correspond more directly to real-world challenges

The panel intended that each of the above artifacts be collected, summarized, and published to help advance the national practice dialogue, resources, and evidence base associated with patient de-duplication.

3 Foundational Concepts for IIS Patient De-duplication

3.1 Objective

The goal of patient-level de-duplication is to correctly match all records related to the same patient even when there are variances in the data used to establish the patient's identity. Matching or linking records relating to the same patient from several or multiple data sources is often required to integrate the information needed to construct an accurate immunization history. Patient matching for record updates and detecting and removing duplicate records that relate to the same patient are processes fundamental to IIS operations. Consequently, the presence of duplicate patient records can lead to inaccuracies and undermine the functionality and credibility of IIS operations.

From an EHR-IIS interoperability standpoint, patient de-duplication can be described in a very simplistic high-level diagram. There are three actions an IIS takes when an EHR sends data into the IIS (Figure 3.1). In general, the IIS can insert a new patient, update an existing patient, or decide that the decision is best made by a human in the context of manual review.

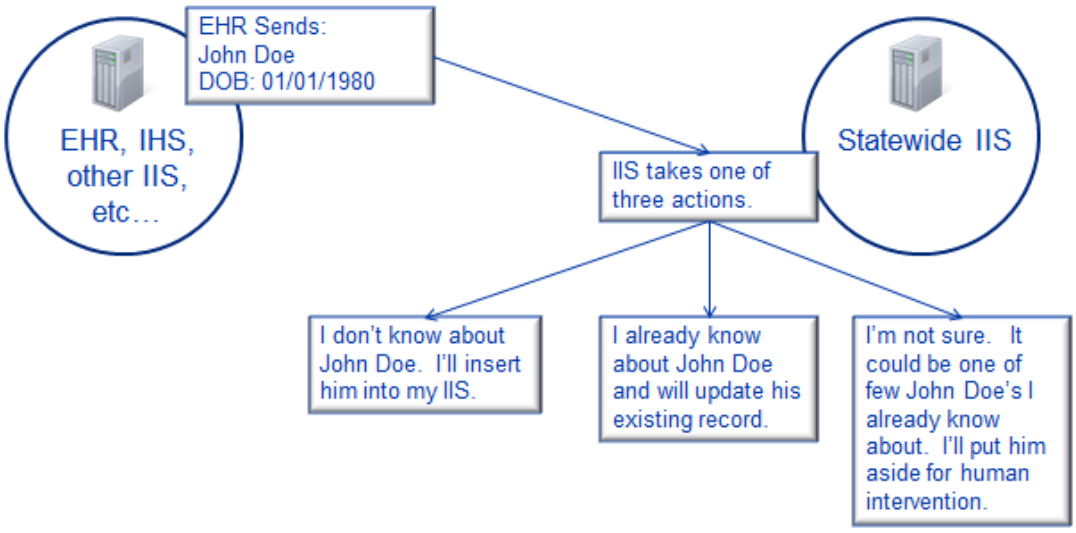


Figure 3.1: Possible Actions Taken When Receiving EHR Data

While the process can be easily depicted, the decision whether to insert, update, or put aside for human review can be very complex. The ability of a system to make these decisions both accurately and on an automated basis determines its overall efficiency.

As IIS mature and grow in size, they receive data from multiple sources (e.g. physicians, hospitals, pharmacies, etc.). Differences in recording data, system design, and data utilization can sometimes result in a patient having more than one record.

The decisions made during patient-level de-duplication affect the forecasting of vaccine administration according to recommendations made by the Advisory Committee on Immunization Practices (ACIP). It is important that inputs into the vaccine forecasting tool include an accurately recorded and consolidated immunization history so that precise and clinically-meaningful immunization decisions can be made. Additionally, correct and complete information on the vaccination history of the patient is essential for providers and analysts so that accurate and up-to-date vaccination history records can be produced.

3.2 Patient De-duplication Processing Scenarios

In general, patient matching and de-duplication processes can occur in three situations: 1) interactively during front-end manual data entry, 2) analysis of data provided through an automated feed from an external immunization provider, and 3) retrospectively through a back-end examination of the database to check for duplications. While patient de-duplication can occur in these three situations (and usually does for larger IIS implementations) variations do occur.

It is believed that in IIS, where large numbers of patient records are coming in from multiple sources, de-duplication processing in all three situations can be beneficial. Front-end manual and automated data processing can act as a gate to prevent bad data from ever entering an IIS system; and routine back-end examination, which maintains a history of adjudication, enables the highest quality accuracy.

The details and complexities associated with these data processing tasks can vary significantly. A good de-duplication process makes good determinations, meaning decisions that are accurate for the jurisdictional data on which it operates. While the standard tasks associated with patient matching and de-duplication can vary, there is conceptual similarity in the processes, regardless of the system being used. However, there is often no single “right answer.”

3.3 Inside the Black Box

For many implementations, except for manual data review, patient matching and de-duplication data processing functions have become information technology (IT)-owned or vendor-owned functions. In some instances, the de-duplication engine has been wholly developed and managed by IIS technical personnel. In other settings, the de-duplication engine has been managed by an HIE, MPI, or other non-IIS specific systems personnel. And, in other cases, the de-duplication engine may be part of a commercial, third-party, or open-source application.

Regardless of how the patient de-duplication processing is administered, the results of the national practice assessment indicated that, to IIS subject matter personnel and IIS administrators, the details of patient de-duplication processing are sometimes only understood on a very high-level, conceptual basis. This is called a “black box” view.

A **black box** is a system which is viewed solely in terms of its input and output characteristics, without any comprehensive knowledge of its internal workings. To improve patient de-duplication methods, it is necessary for all stakeholders to have a greater understanding of the de-duplication black-box operations.

3.4 Five Typical Process Steps

Results of the NPA suggest that IIS administrators and SMEs need to have a greater understanding of the steps and deterministic and probabilistic techniques being used to accomplish patient matching and de-duplication. Moreover, all IIS personnel, both technical and non-technical, need to be participants in an on-going patient de-duplication process improvement dialogue.

As illustrated in *Figure 3.2: Five Major Steps of a Typical Patient De-Duplication Process*, patient de-duplication commonly involves 5 universal processes.

While IIS implementations may use slightly different terminologies and describe the process in slightly different steps, a patient matching and de-duplication process must: accurately find candidate records, block or cluster them into groups of potentially matching records, make decisions based upon business rules and decision logic, and report the outcome of the de-duplication decisions made.

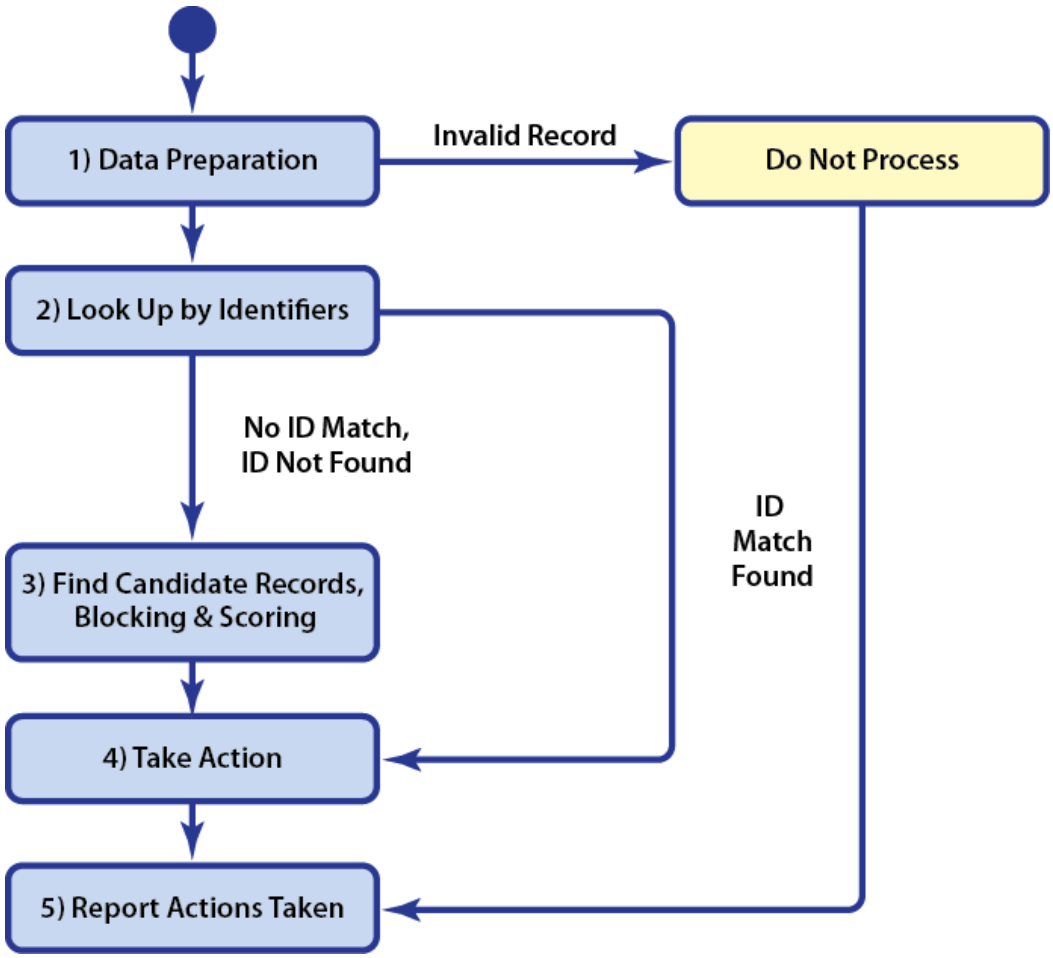


Figure 3.2: Five Major Steps of Typical Patient De-Duplication Process

IIS administrators, SMEs, and technical personnel all need to have a common understanding of the processes that occur within the patient matching and de-duplication black box. Understanding and participating in their set-up and ongoing review can result in significant data quality improvements. This is particularly so when combined with review of audit trails, examination of manual review situations, and use of objective data quality measures.

3.5 Data Preparation

In Step 1, it is recognized that IIS information can come from multiple independent data sources. Preventing duplicate records can be challenging since the format of the data is often different between databases. So, systematic testing of format and content of incoming data within the on-boarding process for a new data source is critical as a first step to ensure quality data. Because there is no way to guarantee consistency between these data sources, discrepancies may arise between what would otherwise be identical data. Data preparation and data cleansing are essential steps in the patient matching and de-duplication processes.

Discrepancies in data can arise from a number of circumstances, including:

- Use of different data entry conventions
- Simple data entry mistakes due to typing or spelling errors
- Extra or omitted data
- Use of non-standard abbreviations
- Misreported or deliberately corrupted data

Data preparation procedures can rearrange and categorize the independent components of any data field including name, address, and other identifying data fields. Invalid data elements may be removed and/or prefix and suffix data may be converted to be stored in a different field. All data may be converted to capital letters or another agreed upon format for storage and comparison purposes. In addition to name data, fields such as date of birth and address may also be checked to see if they are available. Address information may be particularly relevant, if utilized within the IIS. Address information can be in various formats and include abbreviations and other characteristics which make direct comparisons difficult. In order to make address information easier to process by a computer, these data can be prepared, cleansed, standardized, and put in a specific order for comparison.

Data preparation and cleansing algorithms can take several forms. The most common forms include having a reference set of data, such as a dictionary, and performing field edit checks to determine and correct errors and inconsistencies. Selection is followed by data preparation and putting the data into formats where values can be compared.

For patient matching and de-duplication purposes, a minimal set of data is probably also required prior to processing. The data preparation process step can ensure that incoming data meet minimal processing requirements.

Foundational methods for detecting data de-duplication focus initially on using data cleansing methods and making data comparisons as straightforward as possible. More advanced practice considerations will involve reference data and more sophisticated string matching techniques.

Without data preparation, automated de-duplication efforts are handicapped. Poorly designed data preparation procedures can make automated de-duplication problems worse.

3.6 Look-up by Identifiers

3.6.1 Selection of Identifying Data

In Step 2, the selection of the data fields to be used for patient de-duplication is a primary consideration. Different IIS use different data fields to establish patient identity. Common identifying data fields include last name, first name, date of birth, mother's maiden name, address, and selected vaccination data. Based upon business process rules, certain data fields are examined and selected to be used to identify candidate records.

Table 3.6.1: Data Fields Commonly Required in IIS for De-duplication

Variable	Required in IIS %	Required for De-dup %
Patient first name	95.3	69.8
Patient middle name	14.0	4.7
Patient last name	95.3	76.7
Patient alias name (first, middle, last)	2.3	7.0
Patient address	37.2	20.9
Patient phone number	16.3	9.3
Birthing facility	4.7	0.0
Patient Social Security Number	2.3	0.0
Patient birth date	97.7	72.1
Patient sex	67.4	27.9
Patient race	18.6	0.0
Patient ethnicity	18.6	0.0
Patient primary language	0.0	0.0
Patient birth order	16.3	0.0
Patient birth registration number	9.3	4.7
Patient birth state/country	16.3	2.3
Patient Medicaid number	2.3	0.0
Mother's first name	25.6	11.6
Mother's middle name	7.0	4.7
Mother's last name	25.6	9.3
Mother's SSN	0.0	0.0
Father's first name	2.3	0.0
Father's middle name	0.0	0.0
Father's last name	4.7	2.3
Father's SSN	0.0	0.0
Vaccine type	93.0	27.9
Vaccine manufacturer	32.6	2.3
Vaccine dose number	41.9	7.0
Vaccine expiration date	27.9	2.3
Vaccine injection site	23.3	2.3
Vaccination date	93.0	27.9

Table 3.6.1: Data Fields Commonly Required in IIS for De-duplication		
Variable	Required in IIS %	Required for De-dup %
Vaccine lot number	27.9	4.7
Vaccine provider	51.2	11.6
Historical vaccination flag indicator	55.8	9.3

Table 3.6.1: Data Fields Required in IIS for De-Duplication

As previously indicated, the expert panel conducted a national practice assessment study (NPA) of the methods and procedures commonly being utilized for patient de-duplication. This included the data fields used within IIS for patient identification, as well as the data fields commonly utilized for patient de-duplication. The NPA questions around IIS de-duplication variable usage are illustrated above. As indicated in Table 3.6.1, there is commonality around the variables that are used as key identifiers for patient matching and de-duplication purposes. The key notation from the NPA was that there are a large number of data fields that may be optionally leveraged for patient de-duplication processing, depending upon local conditions and circumstances. Data fields that can be optionally utilized include: Social Security Number (SSN), birth order, race, ethnicity, and local medical record numbers or other similar organizational identification numbers.

3.6.2 Specific Unique Identifiers

Experience indicates that the accuracy of a matching algorithm is increased by the number of data elements that can be used in the search. This data can be grouped into a hierarchy; some data and combinations of data provide more value than others. From the IIS perspective, certain data have very high value for patient matching purposes, especially when used in combination with other available data.

If available, specific unique identifiers (e.g., Social Security Number, Medical Record Number, Chart Number, and Birth Certificate Number) can be used to quickly find a match in the IIS and prevent calling the de-duplication engine. When present, these data elements can be of great value in establishing the identity of the patient. In the absence of these types of “high value” identifiers, there may be a number of potential patient matches which need to be systematically examined.

Understanding the content of the patient identifying fields and when they will be used along with other relevant fields is essential to detect and correct data problems and reduce the burden of manual review.

3.7 Find Candidate Records/Blocking and Scoring

In Step 3, the records which could represent duplicates are collected and examined. Most de-duplication approaches follow a two-step process known as **Blocking and Scoring**. As explained

below, both blocking and scoring are generally predicated upon the examination of multiple data fields. Additionally, they can use more advanced data processing approaches to approximate matches between fields. When they do this, a probability or threshold score may be established. This score will determine which records become matches, which records are automatically de-duplicated, and which records will be written to a pending file for manual review.

In general, regardless of the de-duplication method utilized, candidate records need to be identified for consideration. Blocking involves "getting a group of people" within the database that resemble the person of interest. The objective of blocking is to identify the pool of potential candidate records that resemble the patient under consideration. Efficient blocking methods will reduce the number of candidate records from the entire database (millions in some cases) to a smaller number of logical records for evaluation and scoring. Blocking strategies are an important, but sometimes neglected, consideration amongst IIS. Well-developed blocking techniques can improve system accuracy by reducing the overall number of candidate pairs to evaluate. Additionally, poorly-designed blocking strategies can sometimes miss potential matches causing false negatives and fragmenting patient IIS records. It is noted that blocking can affect application performance. Often, it is desirable to have several criteria for blocking, but each criterion can require a separate query to the database. Such queries can be time-consuming when trying to find a patient match. The development of accuracy and performance goals can be a consideration as the number of patients in a database increase.

Scoring involves detailed evaluation of the records found during the blocking phase, assigning them some level of confidence, and returning these results back to the caller of the de-duplication engine. This could include 0, 1, or more matches.

Most matching systems isolate and examine candidate pairs of records to determine their match status. Once each pair is scored and adjudicated, the relationships between similar records within a group can then be evaluated.

Though pair-wise record examinations are currently the most common, it is noted that the aggregated information of records may supply information that is undetected when records are compared in pair-wise fashion. It is also noted that the order in which records are examined may influence the outcome of record comparisons in certain situations.

As discussed later in this section, from a data processing perspective, scoring and matching algorithms fall into four basic categories: 1) single-field comparisons, 2) multi-field matching, 3) rule-based matching, and 4) machine learning. They can be additionally characterized as primarily deterministic (requiring an exact match) or probabilistic (an unclear or close match). The ability to systematically examine data that are not exactly identical and to utilize additional data fields, or the context of the data contained in the records being compared, leverages advanced data processing technologies. These technologies are increasingly mainstream in IIS operations.

3.8 Take Action on Matching and De-duplication Outcomes

In Step 4, the de-duplication processes are executed based upon pre-defined business rules. Based on the record comparison results from the de-duplication engine, the IIS must apply business rules to take the appropriate actions required for a record match, non-match, or the inability of the automated choice to make a confident decision. The IIS "de-duplication engine" and its associated processes will take one or more actions and make a determination regarding the examination of potential duplicate patient records. Depending upon the data source, data quality, and data submission type (user interface or automated), the process flow may differ; however, record pairs will be adjudicated. If enough data or similarity exists that two records are determined to represent the same patient, no new patient record needs to be created. If the data present in the two records is judged to be clearly different, and represent two different individuals, a new patient record is needed. Finally, if comparisons of the available data are inconclusive, the records in question may be written to a pending file and human review and adjudication will be required.

From the standpoint of process improvement, it is important to be able to understand the process and the "reasoning" that the de-duplication used to make its determinations. By creating visibility of the logic and decision processes through audit trails and other logs, SMEs and technical personnel can jointly evaluate how to improve patient de-duplication automation logic and processes.

3.9 Report Actions Taken

The fifth and final step in the de-duplication process is to report actions taken. As suggested above, in this step, the results of de-duplication processing need to be reported so that the outcomes of record comparisons and de-duplication can be recorded and studied to guide system improvements. Even when considered as a black box approach, it becomes apparent that reports on the actions taken by each key process can provide insights into efficiencies and problems.

Working as a team, IIS administrators and their technical support should periodically and systematically examine the manual review files and audit trails associated with patient de-duplication processing, looking for areas of weakness. Such feedback can greatly aid both SME and IT understanding of where process improvements can be made. In addition, this process may also inform the requirements or interventions needed around incoming data. Improving the quality or completeness of source data may vastly improve a system's ability to cleanly de-duplicate patient records.

The ability to resolve and remove duplicate records should be considered an essential business process and involves establishing a number of business rules, defining and developing configurable deterministic and probabilistic business logic, and exposing the outcomes of patient de-duplication processes in such a manner that they can be improved.

Figure 2.9 provides an illustration of the steps that are needed to gain the most from technical and SME business process collaboration. As the diagram implies, a formal, repeatable process is required which analyzes manual review activities and de-duplication process run logs. The purpose of this review is to understand how the de-duplication engine is working (and validate if it

is working as designed), identify any de-duplication trends and weaknesses that can be corrected, improve the accuracy of patient de-duplication results, and reduce the need for manual review activities. In order to improve, objective metrics must be tracked. Additionally, in some instances, it is necessary to provide feedback to immunization data providers.

Efficient and accurate matching of patient de-duplication is of paramount importance to the credibility, efficiency, and usefulness of an IIS. The business process approach to planning, implementing, and documenting patient de-duplication must be a comprehensive consideration. It should be noted that, in many instances, neither the de-duplication engine nor the human SME will have enough information to make a completely accurate determination. Research and access to additional information may be necessary. Understanding and documenting these situations can provide important insights into operational weaknesses.

De-duplication has a number of business process considerations. De-duplication functionality needs to be included to support manual data entry, automatic incoming data feeds, and the retrospective processing of data (cleansing and review of data that has been previously entered into the database). The process associated with the manual review of data records is a key resource/cost consideration. When human reviewers ignore or override decisions made by computer algorithms, these situations may provide important feedback to improve automated operations. Additionally, trends in data originating from provider interfaces must be addressed. There must be mechanisms to communicate with public health partners and data providers. Reports on the actions taken can also include creating objective measures.

Some of the benefits of establishing de-duplication metrics include the ability to:

- Verify that the de-duplication algorithms are operating properly
- Understand why records are pending for manual review
- Report successes and address weaknesses in the de-duplication process
- Make adjustments to the de-duplication process, including match thresholds
- Discuss overall data quality issues and improvement activities
- Better manage resources, people, time, and money
- Develop the business case for additional funding based on measureable data

Greater SME involvement and understanding of the actions taken during patient de-duplication are essential in the lifecycle of de-duplication process management. Accordingly, individual IIS implementations should formalize and document their de-duplication business rules for each step of the matching and de-duplication process.

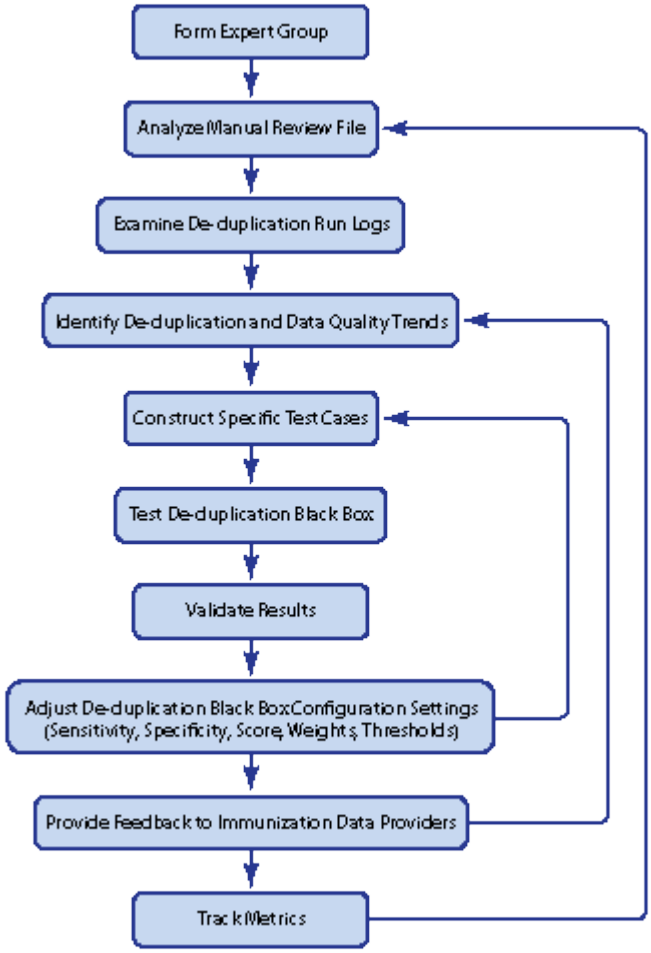


Figure 3.9 Joint Activities Needed For Patient De-duplication Improvement

As shown in Figure 3.9 above, to gain the most from metrics, project SMEs, technical personnel, and data quality evaluation staff should jointly conduct the following activities:

- Review metric reports
- Identify trends, problems, and shortcomings within the de-duplication process by examining the content of manual review files
- Agree upon SME and technical activities that can reduce the burden of manual review processes
- Close the information loop by making sure that there is follow-up by all parties involved in the management of the de-duplication business process

3.10 Classification of Patient De-duplication Approaches

In discussing and evaluating patient matching and de-duplication processes, it is useful to consider where on the continuum IIS processes fall in performing patient matching and de-duplication tasks.

The degree to which de-duplication processes are deterministic or probabilistic is a typical method of characterization.

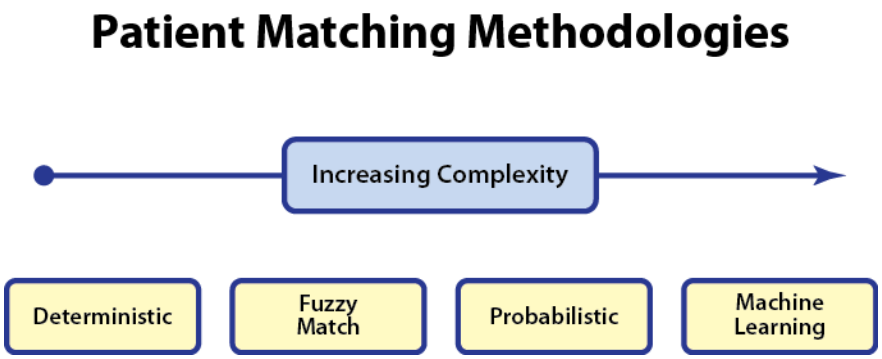


Figure 3.10: Continuum of De-Duplication Methods

On the far left of the continuum shown above in Figure 3.10, the methods are more purely deterministic. Deterministic methods allow little or no room for ambiguities. The data in one record must perfectly match the data contained in another record. As the need to accommodate greater ambiguity increases, more approximate comparisons can be made. Approximates have an established threshold for acceptance.

On the far right of the continuum, hybrid methods that combine elements from deterministic and probabilistic approaches include the examination of more data fields.

While infrequently used in IIS applications, "machine learning," where the computer builds and maintains decision tables that can be refined, represents an advanced method for accomplishing patient de-duplication tasks. Machine learning signifies further sophistication in the efforts to correctly identify and maintain patient records.

In general, as volumes increase and problems become more complex, IIS implementations find the need to progress along the above continuum from deterministic to more probabilistic methods for patient matching and patient de-duplication purposes.

3.10.1 Deterministic

Deterministic de-duplication can be called different names, such as rule-based, exact match, heuristic matching, and other similar terminologies.

In deterministic approaches, programmers work with IIS SMEs to develop matching rules. Deterministic algorithms tend to be straightforward and easy to comprehend; computational requirements are typically minimal. Deterministic models often use exact match or other lightweight field comparators to establish field agreement. Deterministic matching typically identifies matches by defining combinations of field agreement that are believed to reflect matches that are highly likely. The accuracy of deterministic approaches is often greatly dependent on the presence of

"high value" or discriminating identifiers such as a SSN or a local unique identifier such as a medical record number. Deterministic rules tend to rely on the presence of highly specific identifiers and then confirm matches with additional traits.

Deterministic approaches are often utilized in situations where data volumes are relatively low, data entry is predominately manual and well documented, the number of data entry operators is small, and it is possible to enforce data standards.

As the number and types of data sources increase along with variations in data field definitions and content, the efficacy of deterministic methods tends to break down. Deterministic programming can also be used to check for misspellings , transpositions, and create equivalencies (e.g., make "David" equal "Dave", "Richard" equal "Dick", etc.) or make abbreviations equivalent to full spelling, (e.g. "MLK" equals "Martin Luther King"). Accordingly, deterministic examinations of data can become very lengthy and complex as IIS implementations seek to incorporate logic to solve problems involving common data equivalencies.

Experience generally shows that there is a logical upper limit to the effectiveness of deterministic logic. Because deterministic methods rely on accurate and consistent data, they may not generalize well to other healthcare data sources with different data characteristics. This depends upon how the rules are selected and the nature of the variations in data. The national practice experience associated with purely deterministic implementations shows that they eventually hit a ceiling of diminishing marginal returns. The diminishing returns are generally associated with the need to err on the side of failing to match two records for the same patient (also called false negatives).

3.10.2 Probabilistic

In probabilistic approaches, patient matching leverages advanced data. The objective is to have algorithms decide that certain fields match even if they have different values. These algorithms can include looking at the data from different viewpoints such as how similar the strings are in terms of the characters that represent the data, or how similar the data fields may sound if they were to be pronounced. Different data fields may be given different weights that signify their importance of the overall match decision. i.e. not all data fields need to be considered on an equal basis. Low and high comparison thresholds are established and an overall score calculated by the algorithm is compared to the two thresholds. The low and high thresholds establish the probability boundaries for automated decision making. If the score calculated by the algorithm falls between the two thresholds, the two records are held for human review. If the score falls below the low threshold, it is usually not a match. If the score falls above the upper threshold, it is usually a match.

Probabilistic matching algorithms usually start with some standard input data. These data can be entered manually or received from real-time or batch sources. The patient demographic data is compared to the IIS database of patient identities. This comparison usually produces a block of candidate records to which comparisons can be made.

As the fields in these records are compared, they combine these individual field match probabilities to compute an overall likelihood that two different records may represent the same person. The degree of match between two sets of records is determined by comparing against the threshold scores. If the degree of match falls within an established threshold, a match is declared. If the degree of match is close but uncertain, the records are flagged to be reviewed manually.

The low and high thresholds are usually established by programming staff or end-users that have extensive experience determining the dispositions of potentially duplicate records. Ideally, the threshold scores for each situation are agreed upon through active discussion and on-going review with IIS SMEs and technical personnel. The objective is to establish threshold scores which promote accuracy and prevent the false merging of data.

There is risk in assigning an incorrect threshold: if it is set too high, too many false negatives may occur; if too low, too many false positives may occur. Once the threshold parameters are configured, probabilistic decision models enable the declaration of matches between candidate pairs of records which score at established thresholds.

Algorithmic approaches to patient de-duplication require development based upon the characteristics of IIS data sets. As data sets change and evolve, the use of these techniques must also be adjusted. Probabilistic models can produce accurate results within a threshold range. This range can be customized to reflect characteristics of the actual data to be matched. IIS implementations using probabilistic approaches are able to adjust match thresholds to other parameters based upon the specific population characteristics and other features of their patient populations.

As the number of data sources and amount of data increases, probabilistic approaches become increasingly necessary. However, changes will be required to the scoring algorithms. A key part of any probabilistically-based implementation is adjusting the probabilistic parameters to match the patient population within the system’s scope.

This process of “tuning the algorithm” to the IIS population characteristics allows probabilistic algorithms to use a much larger set of matching fields than deterministic algorithms. The probabilistic approach offers a way to trade off the value of different fields in identifying a match.

Parameter and matching threshold customization is typically accomplished through testing and by using human review methods. There is an increasing interest in using automated approaches to review audit trail artifacts to provide useful metadata. However, most IIS systems have not reached this level of sophistication.

Based upon expert panel experiences, there appears to be a predictable life cycle in the development of IIS de-duplication capabilities. As registries are first established and data volumes are relatively low, many implementations have found that simple rules based upon deterministic approaches can be implemented more rapidly and with less technical resources and expertise than

probabilistic methods. As the number of data sources increases and data volumes grow, more sophisticated matching and de-duplication methods are required.

There is an emerging consensus that standard approaches to patient de-duplication are most effective when they combine deterministic and probabilistic methods. It is noted that machine learning approaches also use both underlying methods.

4 Advanced Practice Considerations

Manual record reviews and human processing activities have a number of limitations, are expensive, and may be unsustainable depending upon the volume of records within an IIS. Advanced practice considerations represent a collection of techniques to further automate and improve patient de-duplication processes.

As illustrated in *Figure 4: Enhancing the Five Steps through Advanced Practice Techniques*, shown below, the essential processes that are involved in patient de-duplication can be improved by leveraging more advanced strategies and techniques. These techniques increase the automation of the five step processes and can be combined in ways that improve overall automated decision making and reduce the need for human review. Advanced practice considerations in patient matching and de-duplication processing derive from two broad and somewhat interrelated strategic concepts: 1) patient identity management and 2) overall data quality management.

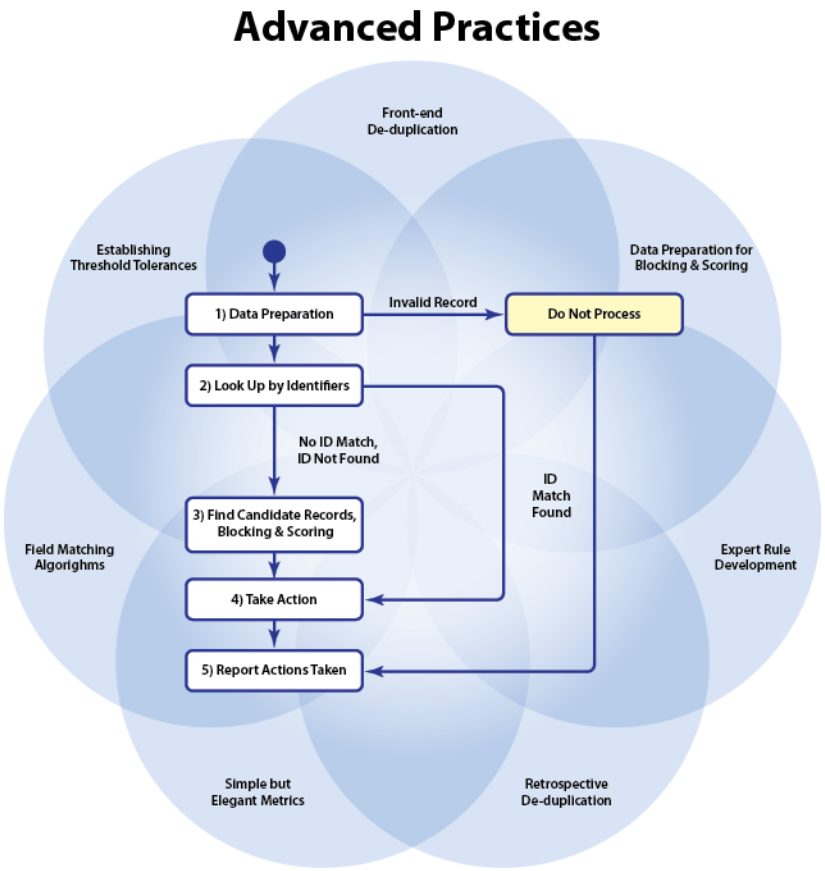


Figure 4: Enhancing the 5 Process Steps through Advanced Practice Techniques

The following sections take a more exacting and advanced look at patient-level de-duplication practices and ways that IIS administration can be improved. Each of the areas summarizes the discussions and consensus viewpoints of the expert panel.

4.1 Front-End Patient De-duplication

Front-end matching is the process of determining if a patient record has already been established for an individual within an IIS. Front-end matching is used in situations where data is being entered manually or where data is originating from incoming data input files or transactions.

The object of front-end matching is to minimize the number of duplicate records that are written to the database. Duplicate IIS records representing the same patient constitute a key source of errors which can erode the accuracy and utility of the overall IIS.

During front-end processing, if a match is found, the demographic data of the existing record, as well as the immunization data contained in the new record, may be used without inserting a new patient record in the database. It is also possible, depending upon the nature of the transaction, that the existing patient demographic information may be updated.

Front-end matching should be considered an IIS best-practice. IIS that support front-end matching allow IIS users to look for potential matches or have automated processes to determine if an incoming record represents a new or existing patient prior to adding a new record into the system.

4.1.1 Manual Data Entry

An effective new patient manual data entry process will force a search or automatically perform a search on the IIS database prior to creating a new patient record even if the user indicates that the patient is new to the system. It should be noted that the nature of this search is generally based upon the data provided, which can vary. The types of searches that are usually done include an ID search, a basic search by name components or date of birth (DOB), and a search employing a majority of the fields used for blocking. This can pull up multiple potential duplicate records for the patient under consideration. Users should be trained on the best search methods supported by their IIS.

Table 4.1.1 illustrates a typical search to determine if a patient record already exists in the Michigan Immunization Community Information System (MICIS). This search is performed prior to adding a new patient into the database.

Table 4.1.1: Example of How To Search For a Child Who Might Exist on MICIS
1. Go to the Child Search screen 2. Click on the Statewide SUI Search radio button

Table 4.1.1: Example of How To Search For a Child Who Might Exist on MICIS
3. Make sure the ISD, District, and School fields do not contain specific values. [All] or [Select] are appropriate. 4. Enter the values in the last name, first name, birth date, and sex fields. If you do not know one of these values, for example the birth date, then change your search to a Partial Info Search. With the partial search, you will not get a closeness score. 5. Enter any of the five fields on the next line that you have: a. Birth City b. Date of First DTP immunization c. Birth Order d. Social Security Number e. Middle Initial 6. Click on the Search button (or just press the Enter key) 7. Examine the results. 8. If no records are on file, then you can enter the child as desired, by pressing the NEW button. 9. If records are present, examine each record to determine if this is truly the same child. If so, then use that record for entry. You may need to review the information in “How to Gain Entry Access to a Child Attending Another District.” 10. If you find what you wanted to know but don’t want to do any entry, make sure you are using MICIS in a professional manner.
Source: Courtesy of MI IIS

Table 4.1.1: Example of How to Search For a Child Who Might Exist in MICIS

There can be many variations in both the data and the edit checks that are performed during manual data entry. When data are submitted through manual data entry, it is preferable that comprehensive edit checks be performed to standardize the data which will be contained in the database. It is also preferable that de-duplication processes be executed to validate the data being submitted and prevent the unnecessary re-entry of data.

From a best practice viewpoint during manual data entry, it is preferable that data be evaluated for completeness, timeliness, and accuracy through online prompting and edit checks, pop-up windows, and other automated mechanisms. On-line help as well as suggestions regarding formatting should also occur during the data entry process. Some systems will show potential duplicate records as data being entered and validated. The ability to rapidly identify a person in a database is a great time saver. Things that slow down the data entry process include instances where it may be necessary to verify the information being presented to the data entry operator.

Training and documentation to assist online data entry practices have also been shown to be effective in improving overall data quality and reducing the instance of duplicates. Additional data

fields may also be available in the application to collect information that the provider and IIS may find useful at a later date.

4.1.2 Real-Time and Batch Electronic Interfaces

The preferred method for immunization provider data submission for most IIS jurisdictions in the United States is an HL7 format. These inputs are accomplished through front-end interfaces to the IIS. The processing of electronic data to the front-end of an IIS system can be accomplished on an on-demand or a scheduled basis.

Despite the increasing standardization of HL7 messages, proprietary flat file formats are still common. Comma-delimited ASCII text files (CSV) are widely used to send batches of data. Depending on the level of automation, providers utilizing this type of option may have less technical capacity and may exhibit poor data quality. For providers using CSV, it may be necessary to pre-screen data more thoroughly and to provide additional pre-processing and cleansing of data prior to de-duplication processing.

Nearly all data contained within interface files will require pre-screening. Business decisions can be made not to utilize certain types of records that may present themselves (e.g. Baby Boy, Baby Girl, BB, BG, non-patient care insurance records).

Establishing well-documented options for providers to submit their data is associated with the success of electronic interfaces and should be considered a best practice.

Routine feedback regarding data quality is also an important facet of data quality management. The feedback given to immunization providers regarding the data quality associated with their data submission helps to manage data quality improvements through active communications and should be considered a necessary and ongoing process. Encouraging providers to review and act upon response files and error messaging is also a critical part of ensuring data quality.

4.2 Retrospective Review

IIS retrospective processing examines the existing records in an IIS database checking for duplicates. Retrospective patient de-duplication (also referred to as “back-end” de-duplication) involves the practice of looking at the data in IIS database and then determining if duplicate patient records exist. The objective of back-end or retrospective review is to identify and resolve duplicate patient records which represent errors in the IIS database. It can also be used to check IIS data quality.

There are numerous techniques that can be used for retrospective review. These techniques include database queries, manual flagging, and “spiders” walking through a database to check for and flag potential duplicate records. Most commonly, a pre-scheduled or on-demand batch

retrospective patient de-duplication process is utilized. Automated retrospective de-duplication processing can be very valuable in preserving overall data quality.

The advantage of a retrospective back-end approach is that the system can automatically examine the entire database and find and resolve large numbers of exact duplicates without any human intervention. Accordingly, retrospective processes can be very valuable for integrated systems where large numbers of records are coming from multiple sources.

The disadvantage of a back-end approach is that it can result in the need for extensive manual record reviews. Researching the records to make a matching determination can be time-consuming and costly.

Once records have been adjudicated through manual review, functionality should exist to retain a record of the adjudication. As a best practice, retrospective de-duplication processing will contain the information needed to not have to reconsider previously adjudicated records.

Threshold and decisions regarding incomplete information as well as other types of decisions need to be formally documented.

Table 4.2: Retrospective Processing	
A record is selected from the IIS database.	Each record in the database is checked against similar blocked records.
Potential matches for this record are found based upon a selection and blocking criteria.	Because these processes can run in the background, their logic can be very extensive. Additionally, different batch processes can be designed to check for specific types of suspected problems.
Pairs of records are evaluated to determine if they are duplicates. Based upon this examination, records may be declared a match and combined, declared a non-match with no other action taken, or be written to a pending file for further human manual review.	Records that have been previously adjudicated can be written to a table in such a manner that they never have to be compared again.

Table 4.2: Steps of Retrospective Processing

4.3 Data Preparation

Data preparation, including data standardization, cleansing, and other forms of preprocessing, are mainly data processing techniques to remove unique formatting. All of these are areas where modest investments in certain common techniques have significant paybacks. The possibility of matching patients and de-duplicating records is greatly enhanced by pre-processing the variables that are to be examined.

The below discussion summarizes the data preparation techniques that should be considered best practices, including name and address cleansing and standardization techniques.

4.3.1 Name Standardization

Each component of a client's name should have appropriate standardization rules.

Table 4.3.1a: First and Middle Name Preparation
• Remove placeholder and "unknown" words, such as "Baby," "Girl," "Unknown," and "None" from the names so that they won't be used in matching. • Recognize and handle middle initials and suffixes that are included in the first name. • Standardize hyphenated names the same way as two-word names. e.g. Mary-Jane is recognized as, or links to, Mary Jane. • Handle nicknames and alternate spellings. e.g. Bob is recognized as, or links to, Robert; Britanni is recognized as, or links to, Brittany. • Remove punctuation. e.g. De'Shawn is recognized as, or links to, De Shawn or DeSHAWN. • Consider using all upper case letters for matching purposes.

Table 4.3.1a: Standards for First and Middle Name Preparation

As indicated in Table 4.3.1b, last name data can specifically benefit from the use of advanced algorithms. These algorithms are discussed in detail in the later sections.

Table 4.3.1b: Last Name Preparation
• Use a phonetic examination algorithm such as Metaphone to produce a phonetic encoding which can be compared to reveal misspellings or other minor differences. • Recognize and handle suffixes and "unknown" values included in the last name. • Handle hyphenated last names the same way as two-part last names. e.g. Smith-Jones and Smith Jones are recognized as, or link to, the same value. • Handle family prefixes in last names. e.g. Mc Donald, McDonald, and Mac Donald are all recognized as, or link to, the same value, as do DeLa Rosa and De La Rosa. • Consider business rules that disallow single character initials for last name that can increase the risk of incorrect merges. • Consider using all upper case letters for matching purposes.

Table 4.3.1b: Benefit of Advanced Algorithms to Prepare Last Name

As summarized in Table 4.3.1c, best practice preparation of address and phone standardization requires a detailed examination of the available address components.

Table 4.3.1c: Address/Phone Data Preparation
• Break down street addresses into smaller pieces. e.g. including house number and suffix, pre-direction and type, street name, suffix type, and unit type and value. A complicated address such as "3301 North Jackson Avenue, Apt 34" would be broken down as follows: House number = 3301 Pre-direction = North Street name = Jackson Suffix type = Avenue Unit type = Apt Unit value = 34 • Standardize the individual values in the smaller pieces to ensure that the maximum match weight is assigned even when the same address is expressed slightly differently. For example, the address components of "3301 N Jackson Ave. Ap 34" would standardize to the same values as "3301 North Jackson Avenue, Apt 34." • Use a phonetic examination algorithm such as Metaphone to produce a phonetic version of the street address to filter out misspellings or other minor differences. • Consider a subscription service of clean addresses which has been geocoded. The address database should be based on United States Postal Service (USPS) or another reliable source to provide reliable address standardization. • Consider using all upper case letters for comparison purposes. • Eliminate "filler" data that can match incorrectly (e.g. 999-999-9999 for phone number).

Table 4.3.1c: Best Practice Preparation of Address and Phone Standardization

In summary, data preparation is an important process. Best practice guidance regarding data preparation best practices includes:

- Standardization of names and addresses
- Standardization of the incoming data for comparison against the IIS data when searching; this will usually produce better results than doing a straight field-by-field match
- Development of algorithms to create sets of likely candidates when comparing standardized data to improve search performance
- Development of functionality to present a list of potential matches for the user for further examination

Technical and non-technical IIS personnel should develop detailed knowledge regarding the behavior of their own IIS implementations. Data preparation can contribute enormously to overall de-duplication process improvement discussions. SMEs and technical personnel should:

- Establish procedures and know the “rules” for data formatting and standardization
- Create documentation that is easy to access and comprehend
- Understand the situations that the de-duplication process cannot handle, such as transposed first and last names
- Know which fields are most important in the de-duplication process
- Provide a human review component to follow the automated component of de-duplication

During operational and data quality review meetings, certain trends and observations should be discussed. Considerations include:

- New variations noted and passed on to technical personnel to incorporate into the standardization process (e.g. MLK as an acceptable abbreviation for Martin Luther King in a street address)
- Culture-specific conventions that will affect the de-duplication process (e.g. family members sharing the same date of birth)

4.4 Blocking

Blocking refers to the process of evaluating relationships among records which could represent the same patient record. Blocking serves an essential pre-screening function. In large IIS databases, the number of candidate pairs to be evaluated can become very large. If blocking is not performed efficiently, the evaluation of large numbers of pairs could hinder the performance of matching and de-duplication processes. Evaluating large numbers of records can cause performance problems, particularly in a real-time matching system. For example, in a large database a single query can sometimes take up to one or two seconds, while evaluating a single record pair can be done in microseconds.

Effective blocking techniques will establish a pattern of agreement but simultaneously accommodate differences in data which can be explored in greater detail based upon the selection of patient identifying data fields. Although many blocking strategies can be described, a common approach is to enforce simple exact-match agreement for different field combinations.

Many IIS block records based upon approximate agreement around 2 to 5 key fields; however, there are many variations. Last names that agree on their first five characters, combined with first names and identical or close dates of birth, can generate a number of candidate records that can be further compared.

Characteristics of ideal blocking fields include high accuracy (few recording errors) and the use of a high number of unique values. Blocking approaches vary for different matching scenarios, depending on the quality of the data being matched and the performance requirements of the IIS. Databases are designed to support blocking functions and blocking functions are designed to take advantage of keyed values.

For example, if the first 5 characters of the last name and the first character of the first name are used for blocking, together they may add a new column to the database and then index that. An index requires more space but significantly reduces query time. Considerations around ideal blocking approaches are typically to optimize the trade-offs between the computational cost of evaluating large numbers of records versus the false negative rates caused by classifying pairs as a non-match. After blocking has been completed, and the candidate records have been selected, more detailed comparisons can proceed.

Examination of blocking strategies can improve patient matching and de-duplication efficiencies. Characteristics of ideal blocking strategies are generally based upon the data fields utilized within a given IIS, along with knowledge of the external data sources. With higher data quality, the ability to standardize agreement of data definitions and contents among external participants improves. While blocking dramatically reduces the time it takes to find a match and makes efficient registry possible, blocking does mean that some potential match combinations will never be considered for matching, even if there is other auxiliary data that would normally help the record become a good match. All of these factors can affect blocking outcomes and the ability to correctly discriminate patient matches from potential record duplications.

4.5 Expert Rule Development

The refinement of patient de-duplication processing can take the form of translated IIS SME experience into effective and computable data processing rules and logic. Expert rule development is central to patient matching and de-duplication improvements. Experts intimately familiar and experienced in patient matching and de-duplication processes (particularly, if developed over a number of years) possess a formidable knowledge base. Such experts can often define rules which can enable automated approaches to correctly discriminate IIS data. The definition of these rules may provide strong circumstantial evidence of patient matches, allow automated approaches to mimic human decision making, and reduce the burden of manual review.

By combining SME expert rules along with high-value data matches and other circumstantial inference techniques (e.g. same names, addresses, telephone numbers, immunization history, etc.), automated decision-making can become more robust. A set of decision rules, including more advanced field-based processing, can then be applied to this new, combined field. Also, if certain unique key fields are present, additional processing may not be required.

As shown below in Table 4.5, some MATCH and NO MATCH field comparisons can also be considered high value. The result of being able or not being able to match data on selected fields can provide a high degree of confidence. The examples in the table are essentially deterministic, not probabilistic, rules; therefore, the aforementioned caveats and limitations of deterministic matching apply.

Table 4.5: Examples of Patient Matching and De-duplication Decision Rules	
Exact match on Social Security Number + exact match on last name or birth date	MATCH. SSN alone or a unique identifier such as medical record number may help short-cut the patient matching process.
Exact match on last name + exact match on birth date + exact match on immunization data	MATCH. The exact match on immunization data, or matches on multiple vaccinations with the same matching dates in addition to other data can create a match
Soundex match on last name + exact match on birth date + exact match on standardized address	MATCH. An approximate match on last name (probably misspelled) along with matches on other data can prevent manual review.
Failure to match on last name + failure to match on birth date	NO MATCH.

Table 4.5: Examples of Patient Matching and De-Duplication Decision Rules

A key consideration for the development of expert-generated, single, and multi-field data comparisons is that they must be clearly specified.

When developing expert rules, the ways in which multi-field matching algorithms combine the results of individual field comparison range from simple logical combinations to complex calculations and should be clearly documented. This documentation can become the basis for record scoring using more advanced field matching algorithms.

4.6 Field Matching Algorithms

Field matching algorithms contain a variety of possibilities and variations including single-field comparisons, multi-field comparisons, rule-based matching, and specific algorithmic approaches to gauge or determine data value similarities. An exhaustive discussion of these approaches is beyond the scope of this publication; however, literature is available which can provide significant insights into effective programming practices.

Single-field comparison algorithms are the simplest and most straightforward to design and implement. Single-field algorithms attempt to find potential matches by quickly comparing individual fields. The contents of the data fields must match exactly. Last name, date of birth, and phone number are all data fields which may lend themselves to a single-field comparison approach.

When data fields are compared, it is not necessary to compare the entire data string to estimate a match. Partial-string comparison techniques are often used. The partial-string comparison category includes string comparison functions that limit comparisons to a specific number of characters e.g. the first three letters of the first name, the first five letters of the last name.

As previously indicated, more advanced algorithms include edit-distance, phonetic, and other types of algorithms that can help determine if the data fields being compared may be different due to misspellings or data entry errors. Additionally, other types of algorithmic techniques includes determining if a numerical or data field value is within a certain tolerance or transposed e.g. the birth dates differ between two records because the month, day, or year of the birth date appear to be transposed. Edit distance techniques can provide an approximation of data entry errors and can alert IIS systems to errors that may have occurred during data entry or transcription. Phonetic distance techniques can provide an approximation of data entry errors on names where the data entry operator is uncertain of spelling or there are spelling differences.

Table 4.6a provides a summary of the most common techniques.

Table 4.6a: Common Algorithms Used to Support Patient De-duplication	
Technique	**Description**
Edit Distance (including Levenshtein, Jaro and Jaro–Winkler distance algorithms)	Edit Distance is a measure of similarity between two strings of data such as last names. In general, edit distance algorithms provide a good approximation of the number of keystroke errors that may have occurred if the two values were supposed to be the same. Edit distance algorithms compute the minimum number of edit operations involving single characters that are required to transform one string to be equivalent or similar to another. Edit operations can include counting such things as inserting, deleting, and replacing characters. The algorithms track and compute these operations and compute the "edit distance" between two strings. The lower the edit distance score, the higher the probability that the data contained in the two strings is equivalent.
Longest Common Substring	The longest common substrings of a set of strings represent a complex routine sometimes found in machine learning systems. The longest common substring represents a complex routine sometimes found in machine learning systems. Like the other string comparison algorithms, this technique provides a way to judge the similarity of two data items by determining the number of character similarities they share.
Soundex (including Metaphone, double Metaphone, and other phonetic comparison algorithms)	Soundex is the basis for many modern phonetic algorithms including Metaphone and double Metaphone. Soundex edit distance comparisons match strings, typically names with different spellings but similar sequences of characters which provide an approximation of similar sounds. Soundex is probably the most widely known of all phonetic algorithms. The reason for this is that Soundex has become a standard feature of some database management languages. The Soundex algorithm essentially indexes the content of string fields (usually names), encoding the similarity of certain English sounds. Using Soundex allows similar names to be matched despite differences in spelling. The Metaphone and updated double Metaphone transformations were developed to improve upon Soundex. It is noted that these algorithms may have limitations for examining non-English names.

Table 4.6a: Common Algorithms Used to Support Patient De-duplication

It should be noted that the above discussion of common algorithms used to support patient de-duplication is not exhaustive and there are many variant techniques in use. For example, the Wisconsin Immunization Registry (WIR) has developed adaptations of Metaphone, double

Metaphone, and other comparison algorithms. In each instance, they may have a higher discriminating power than untailored matching techniques.

By using these different field-value matching algorithms and adapting them to local circumstances, the data contained in different records can be compared. When each data field is compared, the findings of each field comparison can be scored and aggregated into an overall measure of similarity. These comparison scores for each field can be aggregated to indicate the probability that the records are a match.

The typical use of string comparison algorithms is summarized in Table 4.6b below.

String Comparison Summary

Method	Phonetic Misspelling	Typo	Start More Important	Reversed Parts	Missing Part	Tuned for English	Usable for Blocking	Standardized
Soundex	x					x	x	x
Metaphone	x					x	x	x
Double Metaphone	x					x	(x)	x
Jaro-Winkler	x		x			x		
Edit Distance		x						x
Longest Common Substring				x	x			

Table 4.6b: Typical Usage of String Comparison Algorithms

As illustrated in Table 4.6c, in various settings, different algorithmic approaches have been determined to be more or less effective for different types of string comparison problems. In a study made available courtesy of Andrew Borthwick and colleagues at ChoiceMaker Technologies, the various algorithms and their usage are illustrated below (Hall, 2012).

Performance on Selected Name Pairs

Method	Mustafa, Mustapha	Maxwell, Michelle	Carlo, Carlos	Nachanon, Natchanon	Jolanda, Yolanda	Li, Lilian	Yasira, Vasira	jimsmith, smithjim	Jim John Smith, Jim Smith	John Smith, Rob John Smith
Soundex	x	x							x	
Metaphone	x			x						
DBL Metaphone	x				x					
Jaro-Winkler	0.89	0.80	1.0	0.93	0.93	1.0	0.92	0.95	0.83	0.70

Method	Mustafa, Mustapha	Maxwell, Michelle	Carlo, Carlos	Nachanon, Natchanon	Jolanda, Yolanda	Li, Lilian	Yasira, Vasira	jimsmith, smithjim	Jim John Smith, Jim Smith	John Smith, Rob John Smith
Edit Distance	2	4	1	1	1	4	1	6	5	4
LCS	0.71	0.43	1.0	0.75	0.86	0.0	0.83	1.0	1.0	1.0

Graphic adapted from the work of Andrew Borthwick and colleagues at ChoiceMaker Technologies, Inc.

Table 4.6c: Effectiveness of Different Types of String Comparison

A study was performed to determine the accuracies of various combinations of string comparison techniques compared by applying the techniques to pairs of records that were similar to each other (Hall, 2012).

Table 4.6d below summarizes the effectiveness of combining these techniques, as the ChoiceMaker machine learning product does, and reducing the need for human review (Hall, 2012).

The first row shows that without any correlation tests based on approximate string matching (i.e. using only exact agreement between fields), 62% of pairs required human review (Hall, 2012). As string comparison tests were added to the machine-learning algorithm, the required level of human review was reduced.

Results with Different Combos

Soundex	Edit Distance	Jaro	Human Review Percent
			62.8%
X			17.9%
	x		19.7%
			62.1%
		x	12.8%
	x	x	1.7%
X		x	1.5%
X	x	x	2.3%
X	x		3.4%
X	x	x	1.6%

Graphic adapted from the work of Andrew Borthwick and colleagues at ChoiceMaker Technologies, Inc.

Figure 4.6d: Results of Different Combinations of String Comparisons

Table 4.6d also shows that using Soundex in addition to exact matching reduced the percentage of human reviewed pairs to 17.9%. Using Edit Distance reduced the percentage to 19.7%. The

single most effective string comparison technique was JaroWinkler, which reduced the human review percentage to 12.8% (Hall, 2012).

It was noted that the best machine-learning results were obtained by using a combination of string comparison results. By combining Soundex, JaroWinkler, and others; machine learning could obtain the required level of matching accuracy with just 1.5% of pairs requiring human review (Hall, 2012).

A limitation of these observations must be noted. Different techniques may be more or less effective depending upon the definitions and make-up of the data elements contained within each IIS.

Technical and non-technical SMEs should develop an awareness of how to leverage advanced field-match techniques and algorithms to improve patient matching and record de-duplication processes. These implementers should also collaborate with regard to which methods, used alone or in combination, show the greatest efficiencies for individual IIS implementations.

4.7 Establishing Threshold Tolerances

In various circumstances, it may be necessary to establish tolerance thresholds for the comparison of field data.

As shown below in Figure 4.7, threshold scores by their very nature force a trade-off between false-negative and false-positive considerations.

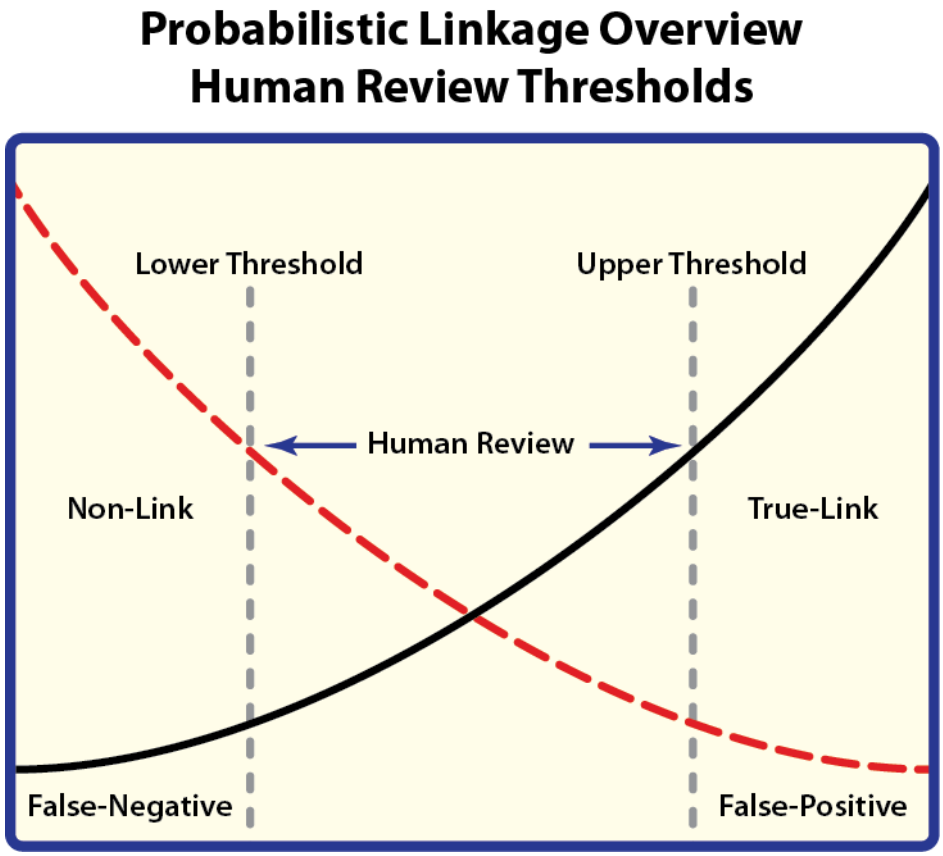


Figure 4.7: Relationship Among Human Review Thresholds and False Negatives and False Positives

When using decision scores and thresholds, it is often useful to characterize the various "regions" that the scores create. Individual implementations usually establish ranges of scores for decision-making purposes. The black curve indicates the number of False-Negative record matches. The red curve shows the number of False-Positive matches. Accordingly, as illustrated, there are trade-offs which much be understood in setting the lower and upper threshold boundaries. These decision scores can be refined to lead to better or more optimal results; however, they can also impact the amount of work which is fully automated or that requires human review.

Records that fall into certain thresholds may require manual review. When manual review occurs, pairs of records are examined manually by IIS SMEs to decide whether they are true matches or not.

There are currently no uniform national guidelines to help immunization practitioners set decision thresholds to improve the quality of IIS data. These decisions are made on a jurisdictional basis. Each individual jurisdiction must set data quality and de-duplication thresholds consistent with the needs of local stakeholders and local constraints and the data that is available.

4.8 Metrics

4.8.1 Five Common Measures of Performance

Measures of patient de-duplication performance are not universally understood or utilized. Additionally, implementing measures in the real world is difficult as there is a lack of "truth data" upon which to objectively make measurements. A universal understanding of metrics can further national practice dialogues in this area.

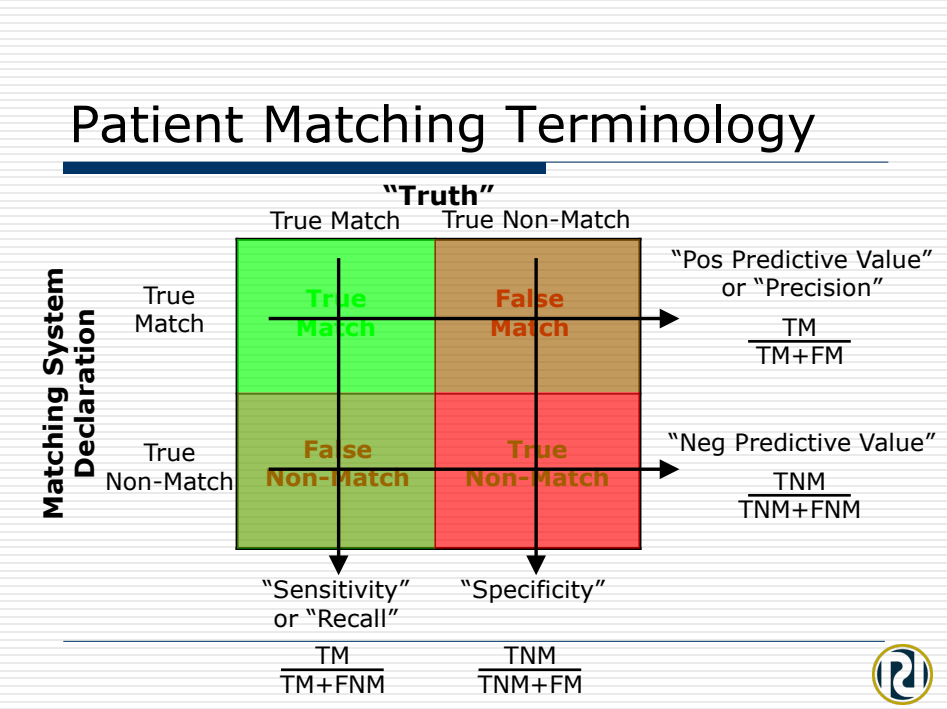
As shown in Table 4.8.1a below, a conceptual understanding exists of five measures that can universally benefit IIS practice efficiency.

Table 3.8.1: Metrics	
Measure	**Algorithm**
Sensitivity – the ability to correctly determine that records are duplicates.	True Positive Matches divided by (True Positive Matches + False Negative Matches)
Specificity- the ability to correctly determine that records that appear to be duplicates are really separate individuals.	True Negative Matches divided by (True Negative Matches + False Positive Matches)
Accuracy- the overall ability of a system to detect patient duplicates.	True Positive Matches + True Negative Matches divided by (True Positive Matches + False Positive Matches + True Negative Matches + False Negative Matches)
Precision- the ability of a system to discriminate true positive patient matches.	True Positive Matches divided by (True Positive Matches + False Positive Matches)
False Positive Rate- the rate at which records are merged in error.	False Positive Matches divided by (True Negative Matches + False Positive Matches)

Table 4.8.1a

The expert panel believes that measures are best used to improve internal jurisdictional IIS operations. To understand how to use these measures for internal improvement, it is necessary to understand the vocabulary and concepts behind each measure.

For IIS public health applications, these measures can be expressed in a familiar 2 X 2 table, as illustrated below in Figure 4.8.1b.



Provided courtesy of Dr. Shaun Grannis, Regenstrief Institute (2012).

Figure 4.8.1b: Measures Expressed in Two by Two Table

4.8.2 Additional Useful Measures

In general, de-duplication functionality can be increased through a formal, routine, and consistent examination of data quality and the content of overall operational parameters. Additional useful measures for understanding IIS operations and improving data quality can include:

- Total, actual new clients added to registry (Unique IDs) – The number of new clients or patients.
- Undetected duplicates added as new records – Number of duplicates that were missed and were added as new clients or patients instead of matched to the other client or patient loaded.
- Duplicate records merged, rejected, or flagged as duplicates – Number of duplicates that were detected as duplicates or possible duplicates. The records that were identified as duplicates were matched to an existing patient and the vaccination record was added, or the record was rejected if it was a duplicate immunization. Some of these records may have been considered as possible duplicates and therefore given a different ID but flagged for manual review. These records may also have been rejected for validation reasons; however, the current analysis tool assumes that if they are not loaded then it is acceptable because two unique IDs were not assigned to the same person.
- Records handled as certain duplicates (merged or rejected) – Number of duplicates that were rejected or merged with the matching records. These were not added as new clients and no manual review is needed.

- Records handled as potential duplicates (flagged) – Number of duplicates that were added as new records but were flagged for manual review to determine whether they really are duplicates.
- Missing records (mistaken for duplicates or rejected) – Records that could not be found in the loaded data. Most likely this would be because they were rejected or merged as duplicates. In a few instances, these records could be missing because they were rejected due to what the registry considers invalid fields.
- Records correctly identified as non-duplicates – Number of unique (non-duplicate) records that were identified and handled as unique records.
- Non-duplicates flagged as possible duplicates – Number of unique records that were flagged for manual review because they were considered possible duplicates.
- Overall score for duplicate record detection – Sensitivity score, calculated as the percent of duplicate records found by the registry out of the actual duplicates in the data.
- Overall score for accuracy in duplicate record determination – Specificity score, calculated as the percent of non-duplicate records found by the registry out of the actual non-duplicates in the data.

4.9 Master Patient Index (MPI)

The emerging role of MPIs in IIS data interoperability is an important developing area. While some models have begun to emerge at the national level and some jurisdictions have begun gaining experience developing the overarching architectures of MPIs, there is currently insufficient evidence-based or practice-based information to definitively guide IIS best practices within the context of MPIs.

A central function of IIS and other types of health information systems is to support the aggregation and exchange of patient data. While initial perspectives on record locator services and MPIs exist, public health registries, including IIS, are at the forefront of patient matching and patient de-duplication problems. There is no documented national consensus on the minimal data that is required for an HIO or HIE to connect. There must be sufficient up-to-date information to support patient matching. Accurate matching of patients across many different systems depends on several critical factors.

The Office of the National Coordinator (ONC) has indicated that having the right patient data, at the right place, starts with accurately capturing and coordinating a patient's identity across multiple disparate organizations (ONC, 2012). If the information presented at the point of care is matched with the wrong patient, it is not only unusable but dangerous for the patient. In the absence of a unique national identification number or some other unified way of identifying people and organizations, the role of master data management (MDM) and MPIs should be expected to evolve significantly in the next decade.

There are a variety of different proprietary and open source approaches that can be used in a master patient index (MPI) to address matching the identities of individual patients that are

scattered across many disparate care settings. These approaches to patient identity management can rely on the use of a unique patient identifier, a voluntary patient identifier, patient biometrics, or an algorithmic matching approach. Each of these approaches has pros and cons; however, consumer rights concerns, financial requirements, politics, and other influencing factors have driven the U.S. healthcare system and data exchange initiatives towards an algorithmic-based set of solutions for cross-system and inter-facility patient identity management (ONC, 2012).

A body of knowledge needs to be formalized which can help further solidify implementations and drive efficiencies acceptable to the national IIS community. The level of accuracy needed for patient records contained in IIS or for EHR-to-IIS data exchanges has not been established. In the absence of a universal patient identifier or national patient identifier, it is unclear if algorithmic approaches to patient matching and patient-level de-duplication can obtain the level of precision needed to sustain widely disparate data interchange operations.

Technical developments and standards such as the **Patient Demographics Query HL7 V3 (PDQV3)** and the **Patient Identifier Cross-referencing HL7 V3 (PIXV3)** provide ways for multiple distributed applications to query a patient information server for a list of patients, based on user-defined search criteria, and retrieve a patient's demographic information directly into the application. The use of HL7 message formats and SOAP-based web services for transport should be expected to continue. Standard messages and the use of SOAP-based web services are well suited for use within an existing IT infrastructure for cross-enterprise data access and exchange. The PIX profile, for example, supports the cross-referencing of patient identifiers from multiple Patient Identifier Domains. These cross-referenced patient identifiers can then be used in various ways to allow authorized health providers to have more complete and accurate patient information—all such developments should be monitored by the IIS national practice community.

Current IIS implementations are successful due to the combination of software and manual review. The volume thresholds at which manual levels of review become unsustainable are undocumented and vary depending upon jurisdictional funding. Without consistent models and standards and the documentation of practice-based experience, the jurisdictional variations in IIS implementations, HIE implementations, and other factors could create barriers to technology adoption and obtaining future healthcare efficiencies embodied in Meaningful Use.

Funding assistance needs to be forthcoming to support deliberate, consistent, and uniform development of standards and models for these areas. Additionally, it is uncertain how the resolution of incoming data and manual entry problems and the review of problematic records within the context of jurisdictional MPI implementations could be conducted and coordinated, if at all.

5 Best Practice Guidance

5.1 General Observations

Based upon the aforementioned efforts, the panel makes the following general best practice-related observations.

1. The expert panel believes that patient-level de-duplication is a significant and important area of ongoing interest to the IIS national practice community. There is only so much that can be accomplished within the scope of work of a single expert panel. The needs of the IIS community are not static. The materials associated with patient-level de-duplication need consistent support and must evolve as experience is gained.

    a. Better mechanisms for sharing and collaboration for on-going patient test case development and discussion are needed. Web-based collaboration tools, beyond a static website, need to be established which will enable efficient leveraging of experience among the national practice community. There are several ways that this can be accomplished. One option to consider is strengthening support to the IIS community through partner organizations, such as AIRA. In connection with this, it is noted that a plan and mechanism for sustainability requires exploration and development. These efforts are beyond the scope of the current expert panel.
    b. Sustained efforts in the areas of EHR-IIS data interoperability and patient de-duplication are required to advance national practice at jurisdictional levels. Meaningful Use implementation places an administrative burden on public health. Very few common tools are available to meet this burden.
    c. To advance interoperability and promote a high level of data quality, additional sustained expert panel work is needed in such areas as establishing minimal standards for external provider data and uniform standards for placeholder data.

2. The expert panel recognizes the desire and need for greater uniformity and standardization of patient de-duplication practices in an IIS. Accordingly, the expert panel has included some of the discussion materials developed in the in-person and virtual meetings as appendices to this final report so that the national practice community can understand and appreciate the scope and breadth of their work. These artifacts include discussions of:

    a. Contextual steps and processes in patient-level de-duplication
    b. Classification and terminology associated with de-duplication software approaches
    c. Desire to evolve evidence-based recommendations to improve de-duplication approaches
    d. Considerations around de-duplication approaches that may enable the greatest productivity approaches
    e. The emerging role of MPIs in relationship to IIS patient identity management.

    Creation of a short summary “best practice” series of Meaningful Use publications should be considered as part of the scope for future projects. It is noted that publications produced by the Modeling of Immunization Registry Operations Workgroup (MIROW) contain relevant content and are consistently referenced by IIS practitioners. It is possible that these materials could be produced from efforts within the IIS national practice community.

3. The information collected in the National Practice Assessment (NPA) needs to be released to the IIS national practice community as a peer-reviewed publication. The NPA helps further the evidence base associated with this important practice area. The NPA revealed that there are wide variations in de-duplication processes, resources, and approaches. This variation is driven by a number of factors including:

    a. Data sources causing patient duplicates
    b. IIS data contents
    c. Patient de-duplication business decisions and practices
    d. Local requirements, needs, and circumstances
    e. Local resources, funding, scope of jurisdictional operations, administrative mandates and mandates of law
    f. Availability and experience of local technical resources

    Understanding these variations in capabilities is important. Consideration should be given to periodically repeating the NPA as a mechanism to characterize the progress and problems in the de-duplication area.

4. There is a need to more formally develop and update the shared vocabulary and terminology associated with de-duplication software approaches:

    a. The literature review and expert panel discussions revealed that there are substantive differences in the terminology used by the different disciplines involved in patient de-duplication processing. While terminology may vary, the contextual steps and processes involved in patient-level de-duplication are generally known and understood within the national practice community.
    b. Based upon various de-duplication software approaches, a more holistic understanding of the contextual steps and processes in patient-level de-duplication can enable practitioners to have greater insight into the systems and procedures that could comprise their roadmap for improvement.
    c. The final report contents are but a first step towards common IIS vocabulary usage. Other mechanisms are needed to integrate and harmonize the unique vocabulary of IIS and health information technology (HIT) and information science academic researchers.

5.2 Best Practice Guidance on IIS Operations

5.2.1 Manual Data Entry and Incoming Data

1. Manual data entry remains an important predominant method of data origination. Despite increasing automation, data origination involves significant human data entry operations. Manual data operations can introduce variations in data, typographical errors, data omissions, and affect overall data quality.

a. Formal documentation regarding manual data entry practices is needed.
b. Greater standardization of IIS data field definitions and manual data entry practices within and beyond IIS implementations may assist de-duplication practices and standards by improving overall data quality.
c. There is a need for greater automation of manual data entry operations, including such interventions as real-time data standardization guidance, prompts for additional data, and real-time examination of the IIS database during data entry for potential duplicate records. All are areas for the more uniform evolution of standardized capabilities which would improve overall data quality; all should be noted by IIS system developers as a largely unmet need. In connection with this, the ability to identify recent status changes may be an important adjunct.
d. During manual data collection and data entry, the identification of life status changes involving marriage, divorce, adoption, guardianship changes, and others may aid in the identification of situations which would create duplicate records or fragment patient histories.

2. The NPA revealed that IIS patient-level de-duplication procedures and practices remain largely un-documented. For documentation, such tools as standard implementation guides, uniform instructions to external data providers regarding IIS data requirements and needs, and procedures for formal reporting and correcting data problems are all useful and necessary activities which can be supported by technical assistance initiatives to the IIS community. In connection with this, the use of web-based materials and/or web-based training is recommended.

    a. Incoming records from external data sources are increasingly the source for patient de-duplication problems. Individual IIS implementations are being required to receive more and more data from external providers. For some IIS jurisdictions, data originating from external providers are now considered among their most problematic data sources and situations.
    b. The NPA revealed that a key source of patient duplicates and poor data quality is data submitted by the immunization provider community. Resources and materials are needed to identify and correct provider training issues and build stronger relationships through routine meetings and feedback regarding overall data quality; continued ongoing support should also be considered in these areas. In addition, standards need to continue to be developed regarding EHR data content and format to ensure high data quality.

3. The expert panel believes that data from external providers needs to be managed systematically and routinely. Incoming records need to be screened and feedback regarding data quality needs to be provided to external data providers.

    a. A mechanism for sharing data quality problems, along with solutions and information which can support greater cross-jurisdictional use and collaboration needs to evolve. Results of the De-duplication National Practice Assessment (NPA) suggest that a number of actions could be taken to benefit IIS data quality.

b. National practice dialogues need to occur concerning best practices around the more uniform management of external data resources including greater standardized guidance regarding the processes for prescreening and rejection of incoming records to reduce manual efforts.
c. There needs to be greater formalization of the procedures used to provide specifications and feedback to external data providers regarding their role in supplying IIS data and the importance of overall data quality.
d. Specifically sustained efforts in the areas of EHR-to-IIS data interoperability and patient de-duplication are required to advance national practice at jurisdictional levels.
e. Data originating from a source that is not approved should not be able to be used in the IIS. In general, incomplete data coming from external data sources do not appear to be utilized for patient care; these external data sources likely do not fully understand the role of the IIS or the importance of the data they are sending. In some instances, it may be useful for contributing immunization providers to be able to run reports. Two types of reports that may be the most useful are Vaccine for Children (VFC) and Assessment reports.
f. Multi-tier approaches are needed. Such approaches would involve reaching out to the broader community rather than just single providers. HMOs, pediatric associations, schools, pharmacies, and other institutions all have an impact on the type of data and its usability.
g. Contributing data partners need to reduce the submission of duplicate patient records; they may need certain capabilities that they currently do not have. Fact sheets, FAQ's, dedicated expert calls, quarterly vendor contact calls, user group exchange webinars, web-based training prior to accessing IIS data, along with recognition and awards for those groups who greatly assist IIS data quality should all be considered.
h. Development of data entry protocols that formally specify what the national data expectations are for IIS are also recommended. Some types of external users and providers require pre-screening and/or pre-processing. Others require read-only access. Information from schools and insurance companies may not be utilized by IIS.
i. Practice recommendations need to be developed that can be communicated back to data sources to help solve these problems on a national scale.

4. Manual data entry and review processes are expensive and time consuming, but necessary activities. Greater automation around manual data review will be needed as data volumes within IIS increase.

    a. An important goal in the manual review of potential duplication records is to correctly adjudicate records. Once records are correctly adjudicated, the IIS needs to recognize these activities on an on-going basis so that redundant record reviews are not conducted.
    b. The ability to perform manual review is currently dependent on the individual experience of the administrator. There needs to be a systematic evaluation of manual review methods. Based upon the NPA, most IIS jurisdictions lack documentation and training around how to best perform these functions.

c. The ability to merge and unmerge patient records in more standardized and graceful ways has emerged as an important area of need from which greater functionality may be required of IIS industry technologists.

5. IIS personnel need to understand the functional differences in their de-duplication approaches for real-time, incoming, and retrospective processing; each may be different and have a different set of strengths and weaknesses. IIS implementers need to pay greater attention to how data interfaces actually operate.
   a. Certain types of records contain placeholder information and may need to be prescreened in order to avoid creating duplicate records.
   b. In the longer term, placeholder information could be standardized by convention on a national scale.

6. When in doubt, practitioners should err on the side of preventing false data merges. The consequences of inappropriately merging the records of two patients are more severe than duplicating a patient's instance in the database.

5.2.2 Retrospective Patient De-duplication

1. Routine, periodic, retrospective examination and de-duplication of IIS patient records should be considered a best practice. In connection with this, certain techniques and practices seem to assist these efforts. As indicated below, it is recognized that both the automated and manual examinations of data records constitute needed practices. Examination of the results of retrospective processing can be an important source for improving overall patient identity management methods as well as improving overall data quality.

2. The results of retrospective processing need to be monitored.

    a. Audit trail and suspense information from retrospective processing can provide important guidance for strengthening the technical and automated approaches of patient matching and de-duplication.
    b. Tracking the number and types of problems associated with patient matching and de-duplication problems provides evidence-based recommendations to improve de-duplication approaches. It can also provide insights into the de-duplication approaches that enable the greatest productivity approaches.
    c. It is believed that retrospective processing could play a much greater role in IIS operations as MPIs are implemented into an immunization system. The emerging role of MPIs is not well documented or understood within the overall IIS national practice community.

5.3 Future Potential Considerations

While beyond the scope of work for this project effort, the expert panel submits the following recommendations for future consideration.

1. The information collected in the NPA needs IIS data contents, funding, and the availability of technical resources. Understanding these variations in capabilities is significant.

2. The IIS community could benefit from additional education and materials around the future roadmaps for EHR-IIS interoperability and IIS Meaningful Use related topics. Additionally, greater awareness, education, and funding are needed to assist the IIS community with participation in the national dialogue on public health Meaningful Use goals, timetables, and evolution.

3. For future efforts, to further promote data interoperability, data quality, and patient identity management efforts, CDC may desire to examine the following ideas:

    a. Strengthening of the working partnership with the American Immunization Registry Association (AIRA). AIRA is viewed by its members as the centralized core of activity for IIS development, standardization, and best practices. A stronger, strategic

relationship with AIRA could be effective in the promotion, advocacy, and dissemination of IIS data exchange standards consistent with national standards. It is believed that joint CDC/AIRA initiatives increase the visibility of IIS public health best practice models including the interoperability and integration of IIS to other health information management system components.

b. Addressing the challenges of patient identity management in real-time query and response environments. Specifically, how can IIS assure quality in a real-time bi-directional environment?
c. Best practices in IIS data standards. A road map is needed to further the jurisdictional mapping of IIS data to NVAC core data elements and NVAC functional standards.

Appendix A - Panel Membership

The patient-level de-duplication expert panel gratefully acknowledges and appreciates the support of the CDC for constituting this panel and providing attention to the important areas of patient-level de-duplication testing and best practice development. The project team appreciates the efforts of the expert panel and their willingness to participate.

The CDC Expert Panelists

Michael Berry is a project manager with HLN Consulting, LLC and has contributed to immunization information systems since 2003. In addition to immunization registries, his work is focused on connecting public health and Health Information Exchange (HIE), standards-based messaging and interoperability architecture, privacy and security in HIE and person-matching technologies for integrated systems. Currently he manages HLN's projects for the Rhode Island Department of Health, and also works as a subject matter expert on the ONC State HIE Cooperative Agreement Technical Assistance Program. berrym@hln.com

Nathan Bunker is a software developer and public health consultant for public and private agencies, focusing specifically on immunization software and data exchange. His work has given him experience with key immunization registry functions, including: immunization recommendation/forecast, HL7 interfacing, data quality analysis, vaccination matching, patient matching, and vaccine barcoding. nathan.bunker@gmail.com

Gerry Bragg, MBA has over 20 years of experience in systems analysis and programming and for the past 15 years, has supported the Michigan Care Improvement Registry (MCIR) as a Senior Systems Developer. He has supported the MCIR system in a variety of capacities, including the development of patient de-duplication/match-merge processes and clinical decision support/immunization forecasting algorithms. Mr. Bragg also specializes in database/SQL performance, scalability, tuning, refactoring, design, technical planning, and configuration management. The system currently supports more than 25,000 users.

Mr. Bragg holds an MBA in Management Information Systems from the University of Minnesota in Minneapolis, Minnesota, and a BA in Accounting from Hillsdale College in Hillsdale, Michigan. He resides with his wife and family in Brighton, Michigan. Gerry.Bragg@altarum.org

Shaun Grannis, MD, MS, FAAFP is a Research Scientist at Regenstrief Institute, Inc. and Associate Professor of Family Medicine, Indiana University (IU) School of Medicine. Dr. Grannis received an Aerospace Engineering degree from the MIT, and underwent post-doctoral training in Medical Informatics/Clinical Research at Regenstrief Institute. He joined IU in 2001. He is a member of World Health Organization (WHO) Collaborating Center for the Design, Application, and Research of Medical Information Systems.

Dr. Grannis completed an analysis of an automated regional electronic laboratory reporting system that revealed substantial increases in the capture rates for diseases of public health. He is project director for an initiative integrating data flows from over 120 hospitals across the state of Indiana for use in public health disease surveillance. This system has received real-time data from hospitals with more than 2 million transactions per year, and has

detected public health outbreaks. As co-chair of the U.S. Health Information Technology Standards Panel (HITSP) Population Health technical work group, he helped lead development of technical Interoperability Specifications.

Dr. Grannis also serves as the Director of the Indiana Center of Excellence in Public Health Informatics, which recognizes that public health practice is driven by a wide variety of data types, data sources, and data management techniques. sgrannis@regenstrief.org

Rick Hall, PhD (Physics) has worked on record matching software since 2003, first at ChoiceMaker Technologies, Inc, and later as an independent consultant. He maintains two open source projects, Open Source ChoiceMaker Technology (http://oscmt.sourceforge.net) and A Data Generator (http://adatagenerator.sourceforge.net). His current clients include the New York City Department of Health and Mental Hygiene, the New York State Department of Education, the New South Wales (Australia) Centre for Health Record Linkage, and the Queensland (Australia) Department of Health.rick@rphall.com

Steve Jarvis is a Data Interface Specialist for the Colorado Department of Public Health and Environment, Immunization Section (CIIS). Since receiving his B.S. degree in Computer Science, he has have been working on various software engineering tasks for the past 25 years, first with group health insurance, then moving into the healthcare field to develop software for clinical management software used by school-based health clinics. In 2003, he was recruited to work with the Colorado IIS and currently manages all aspects of information reported electronically to the CIIS. Steve.Jarvis@state.co.us

Brian Jorgage has worked as a programmer and database developer for over 15 years at various organizations in the Philadelphia area. Over the last several years he has served at the Philadelphia Department of Public Health in support of Philadelphia's city-wide immunization registry. In that capacity, he has processed incoming data files and worked to resolve various data-related issues. His most recent project has been the testing and deployment of a new immunization registry. Brian.Jorgage@phila.gov

Linda Luebchow has been a Data Quality Analyst with the Minnesota Immunization Information Connection (MIIC) for six years, previously holding a similar position with the Minnesota State Registrars' Office, working with the filing of Vital Records. Her day- to-day work is focused on Customer Service and Data Quality, including de-duplication of immunization records. In addition to this she is currently engaged in helping facilities make the switch to HL7 file formats and Real-Time connectivity to qualify for the immunization portion of Meaningful Use.

Linda holds a Bachelor's Degree in Education and is a former Teacher in both Wisconsin and Minnesota. She has two daughters and one grandson (another expected in October) and currently resides with her husband of 35 years in Grand Marais, Minnesota on the North Shore of Lake Superior. linda.luebchow@state.mn.us

Mary Beth Kurilo, MPH, MSW is the Manager of the Oregon ALERT Immunization Information System, or ALERT IIS, and has been with the Oregon State Immunization Program since 2003. Mary Beth worked primarily in health care quality improvement before joining State Public

Health. She completed her graduate work in public health and social work in 2001 at the University of Washington, earning a joint MPH/MSW focusing on Health Administration and Maternal and Child Health topics. She has presented several topics related to immunization information system development and data use at previous National Immunization Conferences, and has co-authored articles in MMWR and other publications. Mary Beth is also the current Board President of the American Immunization Registry Association (AIRA). mary.beth.kurilo@state.or.us

Christie D. Levy is the Branch Director II (Registry Coordinator) with the Mississippi State Department of Health for the Mississippi Immunization Information eXchange (MIIX) registry system in Jackson, Mississippi. Since 2008, she has assisted in the overall function of the Registry ensuring activities are maintained in adherence to the State of Mississippi, CDC, and HIPAA policies and guidelines. Currently, she works as the project liaison between other programs/agency's project managers and District Administrators to develop strategic plans including marketing and recruitment for the MIIX registry system. She is a member of the American Immunization Registry Association (AIRA) which has a mission to "promote the development and implementation of immunization information systems (IIS) as an important tool in preventing and controlling vaccine preventive disease." christie.levy@msdh.state.ms.us

Megan Meldrum is a Research Scientist (epidemiology) with the New York State Department of Health, Division of Epidemiology, Center for Community Health, Bureau of Immunization, New York State Immunization Information System (NYSIIS). Currently she serves NYSIIS as the data exchange liaison, responsible for creating and maintaining data exchange relationships between NYSIIS and multiple data exchange partners including software vendors, private health care providers, and other state agencies; tests modifications and enhancements pertaining to how NYSIIS handles and stores incoming data; analyzes data quality; and participates in the national IIS community as a subject matter expert. mdm06@health.state.ny.us

Chris Pratt is the Technical Manager for the Utah State Immunization Registry, otherwise known as USIIS. He has been with the USIIS Program since 1996. During this time, he has supported the registry through a broad range of IT roles, including Technical Support, Programmer Analyst, Database Administrator, and Technical Manager. Record matching accuracy factors into the duties of each of these IT roles and each role views the problem of record matching from a different perspective. These experiences have given Chris a diversified and firsthand knowledge of the record matching complexities challenging many patient-centric databases today. cpratt@utah.gov

Helen Redfield is a software engineer with over 30 years technical experience supporting health care applications for state government. She has spent the last 16 years supporting the Texas Immunization Registry ImmTrac, and is the registry's technical expert on patient matching, data import, immunization forecasting, and HL7 data exchange. Ms. Redfield holds a BA from the University of Texas at Austin and currently resides in Corvallis, Oregon with her husband. Helen_Redfield@yahoo.com

Bobby J. Sanchez became the Patient Care Training Administrator for Presbyterian Medical Services (PMS) in New Mexico in April, 2012. For the previous 10 years, he worked for the NM

Department of Health in Health Promotion and as the New Mexico Statewide Immunization Information System (NMSIIS) Training Coordinator. While working in NMSIIS he created and made available interactive trainings that improved accuracy and 'cleanliness' of the data and increased provider and user participation. He has served on several national expert panels and was a board member of the American Immunization Registry Association (AIRA). At PMS, he is involved in the training for a statewide electronic system that incorporates, electronic health records, electronic patient management (appointments, schedules, billing), accounts receivable, electronic dental records, behavioral health, e-prescribing, other modules under development and the integration of all these modules throughout the PMS system. bobby_sanchez@pmsnet.org

Cecile Town is a Senior Research Officer assigned from the Centers for Disease Control and Prevention (CDC), National Center for Immunization and Respiratory Diseases, Immunization Services Division to the Indian Health Service (IHS) Immunization Program, IHS Division of Epidemiology and Disease Prevention in Albuquerque, New Mexico. Since 2006, She has served as the IHS Immunization Interface Coordinator, responsible for facilitating and implementing immunization interfaces between RPMS and state IIS. Currently, she works with multidisciplinary groups to ascertain intersystem interoperability; facilitates testing and development of the IHS Immunization Interface Management software (BYIM); actively participate in the national IIS community as a subject matter expert; and provides data exchange support to IHS, tribal, and urban RPMS facilities nationwide. Cecile.Town@ihs.gov

CDC Expert Reviewers

Brandy Altstadter has worked for Scientific Technologies Corporation (STC) for ten years. She has worked in numerous capacities on the immunization registry products, including requirements analysis, development management and managing support. She is currently a Technical Solutions Architect for STC. brandy_altstadter@stchome.com

Noam H. Arzt, PhD, FHIMSS, is president of HLN Consulting, LLC, which has provided HIT services to public health agencies around the country since 1997. Dr. Arzt holds undergraduate, masters and doctoral degrees from the University of Pennsylvania and is active in a number of leading healthcare organizations (HIMSS, PHDSC, AMIA) and standards organizations (HL7, S&I Framework). A frequent speaker at national conferences on healthcare informatics, IIS, and HIE, Dr. Arzt has been supporting the IIS community for nearly 20 years. arzt@hln.com

Justin Ballou
Physicians Computer Company (PCC)
justin@pcc.com

Tammy Clark, RN, BSN, is the Director of the Mississippi Immunization Program. She is currently responsible for planning, implementing, and evaluating all Immunization programmatic activities, including Vaccines for Children (VFC) and the statewide Mississippi Immunization Information eXchange (MIIX). In addition, she also establishes policies and procedures and for the training/developing program of staff nurses, and provides educational in-services for VFC providers, initiating partnerships with key stakeholders. Tammy.clark@msdh.state.ms.us

Robert R. Grenwelge, Jr. is a Public Health Advisor in the National Center for Immunizations and Respiratory Diseases/Immunization Services Division/Program Operations Branch of the Centers for Disease Control. He has been involved in Immunization Information Systems for over 15 years, beginning during his work as the Administration Manager of the Communicable Diseases Division of the Houston Department of Health and Human Services (HDHHS). While there, Robert was involved in the public and private collaboration to develop and implement the Houston/Harris County Immunization Registry (HHCIR). His association with HHCIR continued after he became a CDC Public Health Advisor assigned to Houston's Immunization Bureau. In 2005, Robert relocated to Wyoming as the CDC Public Health Advisor assignee to the Wyoming Department of Health (WDH) and has been involved in the further development and implementation of the Wyoming Immunization Registry (WyIR). robert.grenwelge@wyo.gov

Savonya Jones
Mississippi Patient Information Management System (PIMS)
Savonya.jones@msdh.state.ms.us

John Kellgren is the Lead IT Architect for the District of Columbia Department of Health (DC-DOH), and is currently serving as DC-DOH's EHR-IIS Interoperability Enhancement Project Manager. He previously served as the Project Manager for the DC-DOH's NEDSS Project. A software architect for 25+ years, Mr. Kellgren's background includes experience within multiple commercial industries including construction, distribution, and manufacturing. He has worked on software projects for clients such as Anheuser Busch, Avon, Hercules Construction, Marriott, McDonnell Douglas, and Springfield Remanufacturing. In the late 1980s, he built "Just In Case," a copyrighted software application system designed specifically for health industry case workers. John.kellgren@dc.gov

Tammy LeBeau, BS, has served as the Immunization Registry Coordinator for the South Dakota Department of Health (SD-DOH), Immunization Program since the inception of the program in 1995. Providing the program with VFC/AFIX coordination, she was also developed all of the SD-DOH training materials and conducts yearly trainings for the auditors. As a member of the MIROW expert panel, Ms. LeBeau's contributions to South Dakota's Immunization Information System (SDIIS) garnered the win of AIRA's Center of Excellence Award in 2009 (data use) and 2011 (inventory management). Tammy.LeBeau@state.sd.us

Tammy Lopez
New Mexico Statewide Immunization Information System (NMSIIS)
tammy.lopez@state.nm.us

Thomas Maerz is designer and manager of the Wisconsin Immunization Registry (WIR). An Applications Developer, Computer Electronics Builder and Network Specialist by trade, he has worked with healthcare records and integration with Electronic Medical Record (EMR) systems since 1979 and Vital Records de-duplication of information since 1990. Mr. Maerz has been working with healthcare providers, HMOs, schools and EMR vendors regarding an immunization registry for the state of Wisconsin since 1995.
Thomas.Maerz@dhs.wisconsin.gov

Christopher Qualls
Mississippi Patient Information Management System (PIMS)
Christopher.qualls@msdh.state.ms.us

Laura Rappleye
Michigan Care Improvement Registry (MCIR)
rappleyel@michigan.gov

Lisa Rasmussen has been the Project Leader for ASIIS, the Arizona State Immunization Information System since July 2007. ASIIS reporting is mandated for administered vaccines for children and any pharmacist-administered vaccines within the state system which currently serves over 5 million patients and 50 million vaccines. Holding a Bachelor's Degree in Public Administration and extensive experience in Public Health related databases and systems, Prior to her career in Immunization, Ms. Rasmussen gained additional experience in Maternal and Child Health programs, such as Newborn Metabolic Screening, Health Start, Child Fatality Review, High Risk Perinatal Programs, and Family Planning. lisa.rasmussen@azdhs.gov

Wendy Scharber, RHIT, CTR is founder and president of Registry Widgets, with more than 25 years of experience cancer registration. She is recognized as an international leader in: electronic reporting automated processing of data, and interoperability between public health and eHealth initiatives. She has created and managed electronic reporting systems, data conversion and electronic processing, and rules-based software support systems for cancer registration and specializes in implementing innovative strategies to meet the needs of public health programs. Ms. Scharber is active in several National eHealth initiatives, serving as a bridge between the public health domain and standard setting organizations and implementation efforts. She has authored two profiles within IHE relating to transmission of cancer data from pathology laboratories and from physician offices to the cancer registry. wendy@registrywidgets.com

Lee Taylor, MB is a medical epidemiologist and public health physician in the Centre for Epidemiology and Evidence at the NSW Ministry of Health, Australia. She has managed record linkage at the Ministry of Health since 1994, was instrumental in the establishment of the Centre for Health Record Linkage (CHeReL) in 2006, and is a member of the CHeReL Management Committee. She has extensive experience in the legal, ethical and privacy issues relating to the collection, use, and disclosure of health data for research and management of health services. She is a past chair of the NSW Department of Health Ethics Committee. Dr. Lee also manages data collections that relate to the health of mothers and babies in NSW, and has been responsible for annual reports on the health of mothers and babies in NSW since 1994. ltayl@doh.health.nsw.gov.au

Alexandra Ternier, MPH is a City Research Scientist for the New York City Department of Health and Mental Hygiene / Citywide Immunization Registry. Currently a PhD candidate in Epidemiology, Ms. Ternier's area of expertise includes record matching and de-duplication - probabilistic model development and evaluation.aternier@health.nyc.gov

Karen White, MPH, has worked with the Minnesota Department of Health in the areas of Infectious Disease Epidemiology and Immunizations. She joined the Minnesota Immunization Information Connection (MIIC), Minnesota's statewide immunization information system, in 2002 when the system was first deployed. As an epidemiologist she provides analysis of population-based immunization data to the CDC Sentinel Site IIS project and the EHR-IIS Grant and works as a business analyst for development of additional enhancements of MIIC software. Karen.white@state.mn.us

Northrop Grumman Public Health Contractor Personnel

Frederic Grant, PhD, MPH, MBA, PMP is the Chief Scientist of Northrop Grumman Corporation's Public Health Division which provides advanced IT and business solutions for government and commercial clients. Dr. Grant is an elected member of the Delta Omega Public Health Honor Society. Additionally, he is a Project Management Professional (PMP) and a Certified Data Processor (CDP). Dr. Grant acts as a senior public health advisor to CDC. He is an experienced strategic planner, informatician, public health SME, and facilitator. He has authored numerous publications and industry reports.

Eric Larson is a Senior Information Architect for Northrop Grumman Corporation and is under contract to the CDC Immunization Information System Support Branch. He is currently the lead technical consultant on four EHR-IIS Interoperability Enhancement Projects involving many subject matter experts in the EHR and IIS community. The projects focus on transport layer, HL7, Patient de-duplication and Clinical Decision Support. Previously, Mr. Larson's spent 10 years as an implementer helping several statewide immunization programs implement, maintain and improve their IIS.

Lucretia McKenzie, MPH has worked in the field of healthcare technology for over 14 years. During the past 5 years, Mrs. McKenzie has served as a Business Analyst on various projects for the Centers for Disease Control and Prevention, including supporting software development projects for the Division of HIV/AIDS. She served as a BA for the Clinical Decision Support for Immunization (CDSi) project.

Nina Mitchell currently serves as the Quality Assurance Analyst for the NCIRD De-duplication program and as the Lead Quality Assurance Analyst for CIMS Data Message Brokering for 7 years. Early in her career at Northrop Grumman, she was the lead analyst on numerous successful technical projects from the development to implementation. Over the past 5 years, Ms. Mitchell has supported many CDC programs. Ms. Mitchell provided support in the development of test cases used in application development and complex problem resolution.

Lindsay Ryan is the Project Coordinator for the EHR-IIS Interoperability Enhancement and Clinical Decision Support Projects for the Centers for Disease Control and Prevention. She has been employed with Northrop Grumman for 2 years and has over 12 years of experience in the field of healthcare that spans across multiple focus areas, including reproductive health and medical education. Her prior experience includes coordinating health policy initiatives through state and legislative agencies for both Florida and Georgia, managing the implementation and monitoring of contracts within state/federal government and universities, performing investigative audits for medical records and clinical research and analysis.

Celia Toles is the Northrop Grumman Technical Writer & Editor for EHR-IIS Interoperability Enhancement and Clinical Decision Support for Immunization Projects for the Centers for Disease Control and Prevention. Her prior experience in public health includes database management and project coordination in the Office on Smoking and Health and the Division of Cancer Prevention and Control, for which she participated in multiple standardization and reporting projects for the CDC. Her 17 years of experience in health care also includes positions with WebMD, The Emory Clinic, and the Georgia Institute for Lung Cancer Research.

Jennifer Wain is a Project Manager for Northrop Grumman and has 20 years of experience in project management. As a contractor to the CDC, Ms. Wain currently leads key immunization-related projects in the areas electronic health record/immunization system interoperability and clinical decision support. Her experience includes employment with Accenture and Unisys and supporting clients such as the US State Department, Coca-Cola, and AT&T.

Centers for Disease Control and Prevention (CDC)

Stuart Myerburg, JD, is a Health Scientist, Informatics in the Informatics and Data Analytics Branch (IDAB) at the Centers for Disease Control and Prevention (CDC). Mr. Myerburg has 15 years of experience working in public health. Before coming to the CDC, he served as an Assistant Director of Information Technology at the Rollins School of Public Health. He now leads the EHR-IIS Interoperability and Clinical Decision Support projects in IDAB.

Appendix B - Patient De-duplication Literature Review

Background

Public health is an evidence-based practice. Evidence-based practice requires the examination of findings from systematic and published research and case studies that can be used to inform practice decision making and strategic planning. Additionally, a systematic review of published literature helps inform about the maturity of the state of public health practice, potential needs, and knowledge gaps.

Supporting the work of the CDC Expert Panel was an extensive literature review. This literature review was conducted upon the start of the project to aid with the overall orientation of the expert panel and then continued with updates through panel activities. The expert panel indicated that a literature review could serve several purposes: It could:

- Validate the expert panel's decisions and proposed direction.
- Identify, develop, update, and harmonize vocabulary and common terminology.
- Provide a foundation for the advancement of patient-level de-duplication methods.
- Help determine the state of the peer-reviewed literature and other evidence-based guidance available to IIS patient-level de-duplication practitioners.

Accordingly, the expert panel sought to determine the most authoritative literature resources which could help inform practice, address gaps in knowledge, document and advance best practices, or otherwise facilitate national practice dialogue.

Methodology

In order to support the activities and needs of the expert panel, the literature review supporting this project was conducted in phases aligned to the work of the panel. The literature review focused on peer-reviewed articles, academic studies, special reports, conference proceedings, and relevant government publications published during the last ten years in English which could be used to survey the overall maturity of peer-reviewed literature specific to IIS, including prior national practice assessments. Because of its applied nature, this literature review was not intended to be an exhaustive academic study or meta-review.

Search terms included: Patient matching and patient de-duplication, record linkage, identity management, entity integrity, data de-duplication, and record matching. Related terminologies also used in expanded searches included the relevant topics of duplicate detection, data cleansing, data integration, data integrity, record linking, data de-duplication, name matching, identity uncertainly, entity resolution, fuzzy duplication detection, and entity matching. Additionally, as circumstances warranted, the parameters of the literature review were expanded.

A key criterion for consideration for inclusion in this final report was the degree that the literature content was specifically relevant to approaches and best practices associated with patient-level de-duplication within the context of (IIS) or related situations.

This appendix features highlights of selected articles found. A more inclusive reference list is provided for future use.

Theoretical Roots

The initial idea of a patient record linkage for public health record keeping purposes goes back to Dunn (1946) observations on "Record Linkage" published in the *American Journal of Public Health*. Roughly a decade later, Newcombe (1959) laid the foundation for actually using record linkage theory in actual practice. These beginnings were then formalized by FellegiandSunter (1969).

FellegiandSunter (1969) described how probabilistic decision rules allowed for a comparison of attributes to identify like records. For IIS de-duplication practitioners, the pioneering work of FellegiandSunter (1969), "A Theory for Record Linkage", remains the mathematical foundation for many record linkage applications used by IIS today.

It should be noted that CDC directly or indirectly funded a number of efforts relative to IIS de-duplication studies.

A decade ago, the Salkowitzand Clyde (2003) publication entitled *De-duplication Technology and Practices for Integrated Child-Health Information Systems* provided an exhaustive review of the IIS problems and practices experienced a decade ago.

Salkowitzand Clyde (2003) indicated that de-duplication is the process of removing redundant data from the database, preventing fragmented and duplicated information from getting into the system, and assuring that queries and updates apply to the correct record. Salkowitz and Clyde (2003) noted that duplicate records in any database can cause serious data-quality problems and prohibit an information system from reaching its full potential. The effectiveness of record matching depends on the quality of the data in the individual records (Salkowitzand Clyde, 2003). Additionally, the de-duplication of patient-level records in IIS may have some noteworthy considerations which add to the complexity of de-duplication processes (Salkowitzand Clyde, 2003). These complexities include:

- Data for patients, typically children, come from multiple sources.
- No universal key exists that allows the integrated system to correlate records.
- The presence of alternate identifiers, such as names, is often incomplete or subject to change.

- The original data may contain errors (e.g. keyboarding errors, missing information, etc.).
- There is no standard record structure across systems (i.e., similar fields in the various record structures may have inconsistent meanings).

Salkowitz & Clyde (2003)also conducted one of the first formal national assessments on record matching and de-duplication technologies for child health integrated systems including immunization registries.

The comprehensive and visionary publication entitled *The Unique Records Portfolio, the Public Health Informatics Institute* (PHII, 2006) indicated that information technology was transforming the landscape of health and healthcare. It is noted that most public health information systems are categorical, isolated silos that cannot exchange data; the importance of integration is gaining recognition.

PHII (2006) indicated that de-duplication is the set of processes that link, match, and merge data to integrate or create an integrated view of information for an individual. As a quality assurance measure, de-duplication ranks as a top management issue and a challenge for integration projects whether for private healthcare initiatives or public health. Failure to identify and resolve duplicate records compromises the quality, reliability, and usability of integrated information systems.

De-duplication involves not only software (e.g. matching algorithms), but also organizational (e.g. change management) and people challenges (e.g., staff training). Addressing these challenges is an information systems management responsibility that requires programmatic and technical input, deliberate choices, well-defined activities, and systematic processes. PHII also indicated that to produce high-quality data, an integrated information system must eliminate duplicate records and assign the correct data to each individual (PHII, 2006).

PHII (2006) outlined that public health registries require specific strategies that include:

- A set of policies and procedures guiding the operation of the integration system.
- A technical architecture that supports the policies and procedures.
- An operational plan or set of activities that addresses the core data quality goals of the integration system.
- A method of evaluating the de-duplication processes to determine how effectively and competently duplicate records are being reduced and resolved.

In short, public health leaders must encourage their organizations to present a comprehensive de-duplication strategy that helps validate their overall information system investment (PHII, 2006).

Leonard, Rariden, Beccue, and Shen (2006) conducted one of the few studies specific to public health registries. Their focus was on modifying approaches to patient record de-duplication that would reduce the need for manual review. The approach they used was to apply a probabilistic record linkage method which was tested relative to the traditional method that applied manual review. Using the Illinois State Cancer Registry and a set of immunization registry test cases made

available by the CDC, they tested the effectiveness of certain de-duplication methods. They noted that a modified probabilistic method performed well at identifying true duplicates with sensitivities of 99.8% for the cancer data and 97.7% for the immunization cases compared with 99.8% and 96.4% using the traditional method. However, the modified method was less successful in avoiding false duplicates, with specificities of 99.9% for the cancer data and 89.6% for the immunization cases compared with 100.0% and 94.3% using a more traditional method.

Best Practice Development

The peer reviewed literature relative to developing IIS best practice development is limited. The documents developed by the Modeling of Immunization Registry Operations Workgroup (MIROW) sponsored efforts organized by the American Immunization Registry Association to develop topic-by-topic, IIS best practice, guidebooks for various aspects of immunization registry functionality. While these contain among the best available information, they lack formal publication and peer-reviewed study status.

In 2005, the MIROW Steering Committee conducted an assessment within the immunization registry community to learn which registry functional components were problematic to deploy and could benefit from collective guidance. The outcome of these efforts has been the development of rules and procedures resulting in more accurate and complete representations of vaccination events.

With regard to the literature around the development of IIS best practice guidelines, Williams, Lowery, Lyalin, Lambrecht, Riddick, Sutcliff, and Papadouka (2011) provide an excellent orientation on how to generally develop best practice guidelines within the context of an expert panel specific to IIS. Williams et al. (2011) describe collaborative efforts to develop best practice operational guidelines for IIS including awareness, acceptance, and utilization by the IIS community.

Williams et al. (2011) indicated that business analysis and facilitation techniques were able to be used to support collaboration among IIS stakeholders who analyzed existing practices, brainstormed new approaches, and developed consensus-based recommendations. Accordingly, their guidance was incorporated into the operation of this program.

Williams et al. (2011) also reported that communication of results within the IIS national practice community had important benefits. Based on IIS Annual Report data, from 2007 to 2009 use of the guidelines increased from 46% to 80% of IIS (Williams et al., 2007).

Selected Academic Literature

There is a paucity of peer-reviewed research on de-duplication software approaches specific to immunization information systems. Conversely, much of the literature around the theory of record

(entity) de-duplication area comes from academic information systems and statistical science publications and is centered on techniques that can increase the probability of finding and correcting duplicate entities in any database.

Over a decade ago, Rask et al. (2000) indicated that the medical and public health communities have advocated immunization registries as one tool to achieve national immunization goals. Rask's discussion indicated that participating providers could use registries to consolidate scattered records; to provide an immunization needs assessment for each patient; to provide current immunization recommendations; to promote automated recall of under-immunized children; to document immunizations for schools, preschools, and camps; to help manage vaccine inventories; to provide practice-based immunization coverage assessments; and to calculate Health Plan Employer Data Information Set (HEDIS) reporting requirements for managed care plans.

Rask et al. (2000) noted that although substantial effort has been involved in establishing registries across the nation, only six of the 64 national immunization projects reported active participation from a significant proportion of private providers. Provider submissions were all predicated upon not having to manage redundant, inaccurate, or duplicate records among other factors. They also noted that annual per-patient costs were lowest in the site that used an automated data-entry interface. Of the sites requiring a separate data-entry step, costs were lowest for the site participating in the registry that provided more intensive training and had a higher proportion of the target population entered into the registry.

The research of Miller, Frawley, and Sayward (1999, 2001) in many ways constitutes a starting point for the current evolution of patient-level de-duplication methods. Research of Miller et al., (1999, 2001) was among the first to document the utility of using different demographic data elements for de-duplication purposes. These studies also presented the use of demographic data, patient history data, and the evaluation of record pairs as concepts central to the automated de-duplication of immunization patient records. In two studies, Miller et al., (1999, 2001) explored the utility of utilizing demographic data and vaccination history in the de-duplication of immunization registry patient records. Miller et al. (1999, 2001) noted that duplicate patient records pose a major problem for many immunization registries, as well as many electronic patient record systems.

In their article entitled *Duplicate Record Detection: A Survey*, Elmagarmid, Ipeirotis, and Verykios (2007) noted that in the real world, entities often have two or more representations in databases. Elmagarmid et al., (2007) indicated that the problem of duplicate records has been known and studied for more than five decades.

Duplicate record detection is the process of identifying different or multiple records that refer to one unique real-world entity or object. The goal of entity de-duplication is to identify records in the same or different databases that refer to the same real-world entity, even if the records are not completely identical.

Elmagarmid et al., (2007) indicated that typically the process of de-duplication consists of a number of approaches or stages that make the data comparable and more useable. Duplicate records many times do not share a common key and/or they contain errors that make duplicate

matching a difficult task. Errors in data are introduced from many sources: the result of transcription errors, incomplete information, lack of standard formats, or any combination of these factors. Probabilistic matching techniques were found to be superior for record matching purposes (Elmagarmid et al., 2007).

Christen & Pudjijono (2009) indicated that patient matching methods are typically evaluated using one of two approaches: 1) an algorithm using real-world data can be compared against manually reviewed records, with each potential match determined to be a true match, a true non-match, or an uncertain match; or 2) synthetic data can be used to create an a priori "gold standard" against which the algorithm's performance can be easily measured.

Industry and Government Reports

HISPC (2009) found that there is currently no consensus on patient matching accuracy thresholds or the method used to verify patient identifiers at the time of encounter. HISOC (2009) indicated that accurate matching of patients across different systems, such as hospitals or HIOs, depends on several critical factors. Organizations that connect to the HIO must provide sufficient, up-to-date information to allow for a match. Additionally, patients are under no obligation to inform providers when they move; therefore, a provider may not have the most current demographic information for a patient, making the matching process more difficult. Identifying traits that change over time is another challenge to accurate patient matching HISOC (2009). Each organization employs its own matching algorithm and patient matching methods, resulting in inconsistent results. States are concerned that without consensus on standards, an organization might send faulty data to a health information organization (HIO), which could lead to an incorrect match and potentially affect patient care.

HIMSS (2009) reported the results of several more academically oriented studies summarized by the RAND Corporation. An effective composite key for matching included first name, last name, DOB, last 4 digits of SSN, zip code, and birth year. The data showed that by using a combination of any of the elements except the SSN, the best false positive rate that could be achieved was 1 in 80,000 (using first and last name, DOB, and zip). By adding the SSN element, the false positive rate dropped to 1 in 39,000,000.

The HIMSS (2009) discussion concluded that in order to achieve satisfactory identification with automatic approaches, several different pieces of data are required and good match rates will require a relatively unique data element be added to the query (HIMSS, 2009). Data captured for a probabilistic match must be distinct enough to provide value to the match (HIMSS, 2009).

RAND (2008); HIMSS (2009) published a monograph entitled, *IDENTITY CRISIS: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System.* This report advocated the use of a universal patient identifier (UPI) as a means to improve the accuracy of record linking, but the study also acknowledged the barriers to UPI adoption. Numerous authors have noted that a unique, individual identifier would help simplify patient matching and record de-duplication. A unique identifier is linked to one individual, provides unambiguous identification, is immutable over time with consistent syntax, is simple of concept to implement, and is cost effective

when compared with other solutions. The lack of standardization of data sets, a standard field, or attribute definition and limitations of statistical matching methods contributes to the challenges associated with record linking. RAND (2008) concluded that statistical matching techniques will be required for the foreseeable future.

HISPC (2009) reported that currently there is no consensus on patient matching accuracy thresholds or the method used to verify patient identifiers at the time of encounter. Each organization employs its own matching algorithms and patient matching methods, resulting in inconsistent results (HISPC, 2009).

Master Patient Indexes

The HIMSS (2009b) white paper addressed certain elements of the Nationwide Health Information Network Exchange (NwHIN). The goal of the NwHIN is to allow authorized users to quickly and accurately share health information in an effort to enhance patient safety and improve efficiency of the healthcare system. HIMSS (2009b) notes that in the absence of a UPI, data must be matched based on demographic information and statistic matching techniques based upon establishing a master patient index (MPI).

PHII (2006) noted that the original role of an MPI was to help integrate system operations over broad geographies and systems. To do this requires a number of technology choices. One of the most important choices is the system architecture needed to achieve integration. System architecture is the distribution of key de-duplication roles among the various software components. The distribution dictates the type and level of resources required at the participating programs or a central authority. The overall architecture of the integration system determines what components play the matcher role.

In separate publications Grannis (2008) and Arzt (2006) noted that the matching criteria are built within the MPI solutions or can be positioned above several of the individual components of a jurisdiction. Depending on how they are implemented, these solutions can include criteria- based setting capabilities along with weighting mechanisms to assign a level of match probability. No standard architecture has emerged to support MPI implementations; time, experience, local capacity, and priorities are all potential drivers in the MPI selection process.

Baksi (2009) provided an excellent industry white paper on the integration of MPIs and de-duplication engines from the vantage point of a software architecture roadmap. The objective of Baksi's paper was to clarify the major concepts related to architecture and design of patient identity management software systems for an implementer looking to solve a specific integration problem in the context of an MPI. Baksi (2009) indicated that there must be a clear software architecture roadmap for implementers of patient identity management systems. From the vantage point of IIS, that roadmap is not currently universally clear.

The Office of the National Coordinator for Health Information Technology (ONC) indicated that MPIs are a core element of the infrastructure of HIEs (ONC, 2012). The goal of MPIs is to provide the most accurate and complete information about a patient's health. An MPI identifies all patients in a healthcare setting and provides users unique numbers to organize and extract information

from patient records within the MPI (ONC, 2012). Having the right patient data, at the right place, at the right time is the goal of health information exchange (HIE). This starts with accurately capturing and coordinating a patient's identity across multiple disparate organizations. If the information presented at the point of care is matched with the wrong patient, it is not only unusable; it is also dangerous for the patient. Delivering the right patient information is crucial to realizing the benefits of HIE. In the absence of a unique national identification number or some other unified way of identifying people and organizations, master data management (MDM), much science, and a bit of art, makes this important work possible (ONC, 2012).

References

Arzt, N.H. (2008). Architecture for Person Matching and De-duplication, American Public Health Association (APHA) 136th Annual Meeting, San Diego, CA, October 29, 2008.
Centers for Disease Control and Prevention (CDC), (2011). Progress in Immunization Information System - United States, 2009. MMWR. January 14, 2011. Vol. 60, No 1.,10 - 12.

D'AndreaDuBois, N.S. (1969). A solution to the problem of linking multivariate documents. Journal of the American Statistical Association. 64: 163-174.

Elmagarmid, A.K., Ipeirotis, P.G., Vassilios, S.V., (2007). Duplicate Record Detection: A survey. Retrieved December 11, 2011, from: http://archive.nyu.edu/bitstream/2451/27823/2/CeDER-PP-2007-15.pdf

Fellegi, I.F., Sunter, A.B. (1969). A theory for record linkage. Journal of the American Statistical Association. 64: 1183-1210.

Knuth, D.E. (1973). The Art of Computer Programming: Volume 3, Sorting and Searching. Addison-Wesley. pp. 391–92. ISBN 978-0-201-03803-3.

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

HIMSS. (2009). Patient Identity Integrity. A white paper by the HIMSS patient identity integrity work group. Retrieved December 11, 2011, from:
http://www.himss.org/content/files/PrivacySecurity/PIIWhitePaper.pdf

Miller, P.L., Frawley, S.J., Sayward, F.G. (2001). Exploring the utility of demographic data and vaccination history data in the deduplication of immunization registry patient records. J Biomed Inform. 2001 Feb;34(1):37-50.

National Immunization Program Technical Working Group (NIPTWG), (2001). 2001 Minimal Functional Standards for Registries. Available at: http://www.cdc.gov/vaccines/programs/iis/func-stds.html

Office of the National Coordinator (ONC), (2011). Recommendations on Patient Matching from the Power Team dated 08/17/11. Available from the Office of the National Coordinator for Health Information Technology website.

Office of the National Coordinator (ONC), (2012). Electronic Health Records and Meaningful Use. Available from http://healthit.hhs.gov/portal/server.pt?open=512&objID=2996&mode=2

Public Health Informatics Institute (PHII), (2006) The Unique Records Portfolio. Decatur, GA: Public Healthy Informatics Institute. Clyde &Salkowitz, 2006. Connections. The Unique Records

Portfolio – A guide to resolving duplication records in health information systems, Public Health Informatics Institute.

Salkowitz& Clyde (2003). De-duplication Technology and Practices for Integrated Child-Health Information Systems. Available from the Public Health Informatics Institute.

Williams, W., Lowery, N.E., Lyalin, D., Lambrecht, N., Riddick, S., Sutliff, C., Papadouka, V. (2011). Development and utilization of best practice operational guidelines for immunization information systems. J Public Health ManagPract. 2011 Sep-Oct;17(5):449-56. http://www.ncbi.nlm.nih.gov/pubmed/21788783

Appendix D - Vocabulary

Term	Definition
Black Box / Black Box Approach	A black box is a system which is viewed solely in terms of its input/output and high level processing characteristics, without any real knowledge of its internal workings. A black box approach looks primarily at inputs, outputs, and process results without a detailed consideration of the internal or hidden processes.
Blocking	Process of identifying patients in a database that resemble the record being processed for the purposes of reducing potential candidates matches.
Clinical Data	Data about a current patient encounter including a person's medical history, current condition, or status including data about an immunization event.
De-duplication (Patient-level De-duplication	The process of identifying and consolidating redundant person records in a database.
De-duplication engine/ package/system/service	A software process that identifies and consolidates redundant person records in a database.
Demographic Data	Descriptive data about people that help identify records as being unique.
Deterministic Record Matching	Also known as rules-based record matching, deterministic matching establishes whether two records represent the same person by comparing the fields in the records according to a set of prescribed rules.
False Merge/False Linkage	A mistaken association between two records committed within a database
First Name Placeholder Data	Birth records that have temporary data entered into the first name field rather than proper first name data.
Immunization Information System (IIS)	Confidential, population-based, computerized information systems that collect vaccination data for persons within a defined geographic area.
Incoming Data De-duplication.	The process of identifying patient level duplicates as data is submitted to an IIS from an external data source.
Master Patient Index (MPI)	A single database that is used across defined public health jurisdictions to maintain a consistent, accurate, and current person-identifying data to facilitate patient matching. The objective is to ensure that the identifying data for each patient is represented only once across all systems under the MPI.
Patient Record De-Duplication / Patient-Level De-Duplication	The process of identifying redundant patient records in a database and consolidating or linking duplicate patient records of the same individual.
Batch De-duplication Run (Also see IIS retrospective processing)	A pre-scheduled or on-demand data processing application that identifies, corrects, or reports potential duplicate records.

Term	Definition
Decision Model / Decision Table	Model parameterized utilizing real data
Decision Model / Decision Table Training	A precise way to illustrate how a decision is made. A way of making the decisions behind a complicated process easy to understand, communicate and manage.
Match	A record pair that is declared as reflecting the same patient or same entity.
Non-Match	A record pair that is declared as not reflecting the same patient or same entity
False Negative Match	A record pair that has been erroneously declared as a non-match.
False Positive Match	A record pair that has been erroneously declared a match.
Firing pattern	A set of matching clues and rules that exercised concurrently when the record pairs being evaluated exhibit specific characteristics.
Golden Record	A record that is believed to be accurate and correct in every aspect.
Indeterminate / Possible Match	A record pair that could not be ruled as either a match or a non-match by an algorithm.
Manual De-duplication	The human process where judgments are made to determine if similar individual records represent the same or different individuals.
Match Threshold	A score or numeric representation regarding the degree to which two records match.
Merge Clusters / Merge Groups (also see blocking)	Groups of records potentially belonging to the same patient
Name and Field Matching Algorithms {"Field comparator methods", "string comparator methods"?}	Standardized computer coding methods that have reproducible results and involve techniques or methods to detect such things as spelling variations, phonetic variations, double name, hyphenated names, double first names, alternative first names, and so forth.
Non-Match	A decision that two records do not represent the same patient.
NVAC Core Data Elements	Required and optional data elements representing the fundamental attributes necessary for identifying individuals and for describing immunization events as approved by the National Vaccine Advisory Committee (NVAC).
Opt-In State / Opt-Out State	Attribute of state privacy/security law that determines how patient records can be used by an IIS. A state can choose to opt-out of the registry at any time. Opt-in/opt-out rules and requirements vary across states and can vary across other legal jurisdictions.
Patient	A person whose immunization status is tracked by the IIS.
Patient Record	A database entry that tracks the immunization status for a person.
Placeholder Data	Records that have a placeholder in a data field often because of required fields that are missing or unknown; variance among the placeholder data is high.
Population-Based Data	Data populated with birth records or data limited to a particular jurisdiction, such as a state.

Term	Definition
Potential Match Flags	Records that are believed to represent potential duplicate patients can be “flagged” or tagged as possible duplicates for further evaluation.
Probabilistic Record Linkage / Record Scoring / Weight-Based Approach	A type of record linkage, also known as “fuzzy matching” that considers a wider range of potential identifiers than deterministic record linkage. Weights are computed for each identifier. The weights are used to calculate the probability that two given records refer to the same entity. Record pairs with probabilities above a certain threshold are considered to be matches.
Probable Match	A possible duplicate patient record.
Record Pair	A logical construct that represents the examination of two records to determine if they match or do not match.
Record Linking / Linkage	The term record linkage is used in two distinct ways. First, it can refer to the process of joining a record from one data source to the record of second data source when both records describe the same person. Second, record linkage can refer to a technique that is sometimes used in patient-level de-duplication to create a single entry for a person from multiple duplicate records.
Record Merging	A technique that is sometimes used in patient-level de-duplication to create a single entry for a person from multiple duplicate records.
Retrospective De-Duplication / Batch De-duplication	IIS retrospective processing examines the existing records in an IIS database checking for duplicates.
Scoring	Process of evaluating blocked records and assigning a measure of how closely they compare to the incoming record.
Sensitivity	A measure of how well a system performs at recognizing duplicate records.
Sensitivity Score	Percentage of duplicate records found by a de-duplication process, out of the actual duplicates in the data.
Sequential Approach	Process that applies a set of decision rules for comparisons of variables in patient records under consideration. The rules evaluate individual variables as well as distinctive combinations of variables. This approach provides the basis for a deterministic approach to de-duplication of records. The sequential approach can be made more efficient by testing variables or combinations of variables with the most discrimination power. Rules-based approaches share some commonality but are typically tailored to local circumstances and issues.
Specificity	Value reflecting how accurate the duplicate record detection is by measuring the rate at which non-duplicate records are misidentified.
Specificity Score	Percentage of non-duplicate patient records found, out of the actual non-duplicates in the data.
Standardized Computer Coding Method	Techniques or methods within name and field matching algorithms used to minimize such things as spelling variations, phonetic variations, double name, hyphenated names, double first names, alternative first names, and so forth.

Term	Definition
Test Case	Data created specifically for the purpose of exercising the ability of a process to perform correctly.
Threshold Score	A score or numeric threshold at which one concludes that a record is a duplicate, is not a duplicate, or may require manual review.
Twin-ness	Ability of a de-duplication process to correctly identify records pertaining to multiple simultaneous births to the same mother.
Undetected Duplicates	Total number of duplicate records that were added as new patients instead of being properly matched to an existing patient.

Appendix E – Document Management

Date	Comments	Version #
06/25/2013	Initial Version	1.0
02/04/2025	Updated CDC Branch from IISSB to IDAB Replaced the term gender with sex	1.1