Vaccination Level Deduplication in Immunization Information Systems

Recommendations
of the American Immunization Registry Association (AIRA)
Modeling of Immunization Registry Operations Workgroup (MIROW)

December 7, 2006

Suggested citation:
AIRA Modeling of Immunization Registry Operations Workgroup (eds). Vaccination level deduplication in Immunization Information Systems. Atlanta, GA: American Immunization Registry Association. December, 2006.

AIRA Modeling of Immunization Registry Operations Workgroup's roster

Chair:

*** Warren Williams, MPH**
Centers for Disease Control and Prevention
National Center for Immunization and Respiratory Diseases
Lead Public Health Analyst
MS E-62, 1600 Clifton Road
Atlanta, GA 30333
Phone: (404) 639-8867
E-mail: wxw4@cdc.gov

Participants:

Bridget Ahrens, MPH
Vermont Department of Health
Immunization Registry Manager
108 Cherry Street
Burlington, VT 05402
Phone: (802) 951-4094
Email: bahrens@vdh.state.vt.us

Brandy Alstadter
Scientific Technologies Corporation
Department Manager
65 E. Weldon Avenue, Suite 110
Phoenix, AZ 85012
Phone: (602) 241-1502
Email: Brandy_Altstadter@stchome.com

Diana Bartlett, MPH
Centers for Disease Control and Prevention
National Center for Immunization and Respiratory Diseases
Health Scientist
MS E-62, 1600 Clifton Road
Atlanta, GA 30333
Phone: (404)639.1813
Email: zxd5@cdc.gov

*** Barbara Canavan**
Oregon Immunization ALERT, Director
800 NE Oregon Street, Suite 370
Portland, OR 97232
Phone: (971) 673-0280
Email: Barbara.C.Canavan@state.or.us

Anne Cordon, MPH, CHES
San Diego Regional Immunization Registry (SDIR), County of San Diego HHSA
P.O. Box 85222, P511B
San Diego, CA 92186-5222
Phone: (619) 692-8403
Email: Anne.Cordon@sdcounty.ca.gov

*** Mike Flynn, MA**
New York State Department of Health
Immunization Registry Manager
Corning Tower, Empire State Plaza
Albany, NY 12237
Phone: (518) 473.4437
Email: mkf06@health.state.ny.us

Sandra Frawley, PhD
Yale University
Associate Research Scientist
PO Box 208009
New Haven, CT 06520-8009
Phone: (203) 737-5812
Email: sandra.frawley@yale.edu

Ruth Gubernick, MPH
RSG Consulting
Public Health Consultant
5 Woodbury Drive
Cherry Hill, NJ 08003
Phone: (856) 751-0115
Email: gubernrs@hln.com

*** Therese Hoyle, BSHE**
Michigan Department of Community Health
MCIR Coordinator
201 Townsend Street
Lansing, MI 48909
Phone: (517) 335-9340
Email: HoyleT@michigan.gov

Laura Korten, MPH
Michigan Public Health Institute
Interactive Solutions Group
MCIR Project Coordinator
2440 Woodlake Circle, Suite 190
Okemos, MI 48864
Phone: (517) 324-7395
Email: lkorten@mphi.org

Marlene Lugg, DrPH
Southern California Kaiser Permanente
Immunization Coordinator/Project Manager
13652 Cantara Street, South 1 Bldg.
Panorama City, CA 91402
Phone: (818) 375-3078
Email: Marlene.M.Lugg@kp.org

*** David Lyalin, PhD**
CDC Information Technology Support Contract
Consultant - Business Analyst
3375 NE Expressway
Koger Center/Harvard Building
Atlanta, GA 30341
Phone: (678) 530-3583
E-mail: dlyalin@cdc.gov

Thomas Maerz
Wisconsin Department of Health and Family Services
WIR Systems Manager
1 West Wison, Room 318
Madison, WI 53701-2659
Phone: (608) 261-6755
Email: maerztr@dhfs.state.wi.us

Vicky Papadouka, PhD, MPH
NYC Department of Health and Mental Hygiene
Director of Research and Evaluation
2 Lafayette Street, 19th Floor
New York, NY 10007
Phone: (212) 676- 2331
Email: VPAPADOU@health.nyc.gov

Sherry Riddick, RN, MPH
CHILD Profile (WA State Immunization Registry)
Immunization Registry Operations Manager
999 3rd Avenue, Suite 900
Seattle, WA 98104
Phone: (206)205-4139
Email: sherry.riddick @ metrokc.gov

Steve Rose
Software Partners
Project Manager
904 Second Street
Encinitas, CA 92024
Phone: (760) 944-8436, ext. 13
Email: steve@swpartners.com

*** Susan Salkowitz, MGA**
Salkowitz Associates, LLC
Principal HIS Consultant
Two Independence Place, #1601
233 S. 6th St.
Philadelphia PA 19106
Phone: (215) 351-7887
Email: salkowit@hln.com

Rob Savage
EDS
Senior Business Services Analyst
10 E Doty, Suite 200
Madison, WI 53703
Phone: (608) 221-4746, ext. 3641
Email: savagrb@dhfs.state.wi.us

Emily Peterson-Stauffer
Minnesota Department of Health
IT Supervisor/MIIC Director
625 N. Robert Street
St. Paul, MN 55164-0975
Phone: (651)201-5546
Email: emily.peterson-stauffer@health.state.mn.us

*** Cynthia Sutliff**
American Immunization Registry Association (AIRA)
Executive Director
C/o Citywide Immunization Registry, NYCDOHMH
125 Worth Street, CN 64R
New York, NY 10013
Phone: (212) 676-2325
E-mail: csutliff@health.nyc.gov

Kent Ware
Ohio Department of Health
Vaccine Warehouse/Registry Administrator
35 E Chestnut St
Columbus, OH 43215
Phone: (614) 466-4643, ext. 229
Email: kware@odh.ohio.gov

Laura Zimmerman, MPH
Centers for Disease Control and Prevention
National Center for Immunization and Respiratory Diseases
Health Scientist
MS E-62, 1600 Clifton Road
Atlanta, GA 30333
Phone: (404) 639-8690
Email: laz5@cdc.gov

* Denotes contributing members of the MIROW that also served in the MIROW Steering Committee.

Acknowledgements

Members of the Modeling of Immunization Registry Operations Workgroup (MIROW) appreciate and acknowledge the following

- External reviewers of the recommendations document for their willingness to read and constructively critique the workgroup's materials; their efforts benefited this document significantly:
  Dr. Noam Arzt, HLN Consulting, LLC, San Diego, California,
  Kevin Davidson, QS Technologies, Inc., Greenville, South Carolina,
  Oliver Eng, EDS, Madison, Wisconsin,
  Katie Reed, Partners In Health Systems LLC, De Witt, New York,
  Sandra Schulthies, Utah Department of Health, Salt Lake City, Utah,
  Cherie Thomas, Trey Industries, Inc., Locust Grove, Virginia.

- Valuable contributions of the AIRA staff—**Angelica Velez** and **Ina Kichen**—in organization of modeling sessions.

- Effective support provided during the modeling sessions in Washington, D.C. (May 8–10, 2006) by the facilitation team from Advanced Strategies, Inc. —**Steve Farrell, Kahil Branton,** and **Frances Caggiano-Swenson**.

- Contributions of **Wisconsin** and **Michigan immunization registries** in providing materials on vaccination level deduplication that helped in framing the topic and preparations for the modeling sessions.

- **Elaine Lowery**, JD, MSPH, Program Director, Colorado Immunization Information System, University of Colorado Health Sciences Center who joined the vaccination level deduplication project at its late stages as a co-chair of the MIROW.

About MIROW vaccination level deduplication project

The Modeling of Immunization Registry Operations Workgroup (MIROW) of the American Immunization Registry Association (AIRA) was formed to develop a topic-by-topic Best Practice guidebook for various aspects of immunization information systems (IIS) functionality.

In April 2005, the MIROW Steering Committee conducted an assessment within the immunization registry community to learn which registry functional components were problematic to use and could benefit from collective guidance. The first topic selected for analysis and development of best practice recommendations was the management of the "Moved or Gone Elsewhere" (MOGE) status of patients and other patient immunization designations, had been selected for analysis and best practice recommendations. That report has been produced (in June –December 2005) and is available at the AIRA web site at http://www.immregistries.org/docs/MIROW_MOGE_Chapter_Final_122005_rev1.doc .

The current report represents MIROW efforts to develop best practices recommendations for the second topic chosen, vaccination level deduplication.

Table of Contents

Introduction ……..... 12

Scope: vaccination level deduplication……. 12

Why is this important? ……... 12

Expected benefits ……... …….. 12

Intended audience ……. 13

Intended use …….. 13

Workgroup approach …… 13

National initiatives of importance to IIS and vaccination level deduplication strategies 14

How these materials are organized …… 15

1 ***Executive Summary*** …….. 19

Recommendations: chapters 2 - 6

2 ***Process Overview***……..... 20

3 ***Selection Phase*** …….. 22

Discussion of the recommended size for the date window ……..... 23

Illustrative scenarios …….. 25

4 ***Evaluation Phase*** ……... 27

Variables …… 27

Two approaches for the Evaluation phase ……. 29

Distinctive combinations of variables for the Evaluation phase ……..... 32

Principles and Business rules……. 34

Weights and aggregated score ……..... 37

Illustrative scenarios ….……. 39

Clear-cut illustrative scenarios …….. 41

Complex illustrative scenarios ……..... 48

5 ***Resolution Phase*** ……..... 54

Selecting the best record and creating a consolidated record …… 54

Discussion of level of confidence in vaccination data ….……. 56

Variables ……..…….. 57

Sequential (deterministic) rules-based approach …..…… ……..... 58

Weights-based (probabilistic) approach …… 63

Ownership for the immunization record …… 65

Management of duplicate records …….. 65

Illustrative scenarios …….. 66

6 ***Additional Miscellaneous Recommendations***……... 72

Barriers for deduplication and ways to address those barriers ……..... 72

Provider-related recommendations …….. 73

The role of invalid doses (e.g., due to improper storage) in vaccination level deduplication ……..... 74

Manual review ……... 75

False matches (false positives) and false non-matches (false negatives) ..…… 76

7 ***Conclusions*** ……..... 79

References…….……. 80

Appendix A: Domain Model ……. 81

Domain diagram for vaccination level deduplication …… 83

Entities and attributes (terms and definitions) ……... 84

Appendix B: Merging Data From Duplicate Records …… 97

Issue in brief …….. 97

Background …….. 97

Example ……. 97

Example - Best record approach ……....…… 99

Example - Consolidated record approach ……. 100

Can consolidation of information for a vaccination event be justified? …….……... 101

Recommendations ……... ……. 102

Illustrations

Map: Recommendations in context of the vaccination level deduplication process …… 18

Figure 1. Process overview diagram for vaccination level deduplication in IIS …….…... 21

Table 1. Variables for the Selection phase ……..... 22

Table 2. Principles and business rules for the Selection phase ….…….. 22

Figure 2. Selection phase overview ……... 24

Decision tables with scenarios for the Selection phase ...…….. 25

Table 3. Variables for the Evaluation phase …….. 28

Figure 3. Evaluation phase overview …… 30

Figure 4. Approaches for the Evaluation phase of the process ……. 31

Table 4. Distinctive combinations of variables for the Evaluation phase ……. 33

Table 5. Principles and business rules for the Evaluation phase ……. 34

Table 6. Weights for key variables: Evaluation phase …….. 37

Table 7. Example of weights assignment for key variables: Evaluation phase ……. 39

Decision tables with scenarios for the Evaluation phase ...…… 41

Table 8. Weights-based approach example: overview of illustrative scenarios for the Evaluation phase …… 53

Figure 5. Resolution phase overview—a two-step approach …… 55

Table 9. Matrix of confidence levels for vaccination data: based on the attributes of "Vaccination Event Submission" entity ……..... 56

Table 10. Variables for the Resolution phase (selecting the best record) …….. 57

Table 11. Principles and business rules for the Resolution phase (Step 1. Selecting the best record) …… 58

Figure 6. Resolution phase: sequential implementation of business rules to select the best record ……. 60

Table 12. Principles and business rules for the Resolution phase (Step 2. Creating a consolidated record) …….. 61

Figure 7. Resolution phase: implementation of business rules to create a consolidated record - merging information for a single variable ……. 62

Table 13. Weights for key variables: Resolution phase …… 63

Table 14. Possible table of weights (used for illustrative scenarios) ……. 64

Decision tables with scenarios for the Resolution phase ...…… 66

Table 15. Matrix of evaluation outcomes for two records (23 days apart or less) …… 78

Figure A-1. Domain diagram for vaccination level deduplication ……... 83

Table A-1. Entities and attributes (terms and definitions) …...……. 84

Figure B-1. Illustration / example for the consolidation/composite record issue …….. 98

Table B-1. Pros and cons for the best record approach …… 99

Figure B-2. Best record approach - for clinical purposes …….. 99

Table B-2. Pros and cons for consolidated record approach …… 100

Figure B-3. Consolidated record approach for public health purposes and clinical purposes 100

Introduction

Scope: Vaccination level deduplication

Deduplication of immunization records is a two-fold problem that includes deduplication at the demographic/patient level (e.g. two records describe the same patient) and deduplication at the vaccination event level (e.g. two records describe the same immunization). While the former is a very common generic task for a variety of information systems, including systems outside of public health, the latter is a domain-specific task for the immunization information field.

As IIS mature and grow in size, they likely receive data from several sources (*e.g.*, the physician, the insurance billing system, etc.) sometimes resulting in a patient having more than 1 vaccination record for a single vaccination event. IIS must then decide how to identify and how to deal with duplicate vaccination entries.

The major focus of this project is on the issues of vaccination level deduplication for lifetime immunization records. This includes development of rules and procedures that result in accurate and complete representation of a vaccination event when compiling vaccination history from multiple data sources. The project also includes development of a consistent set of terms and definitions. This analysis is not focused on demographic/patient level deduplication rules and procedures or clinical decisions related to the immunization schedule.

Why it is important?

Inconsistency among immunization information systems (IIS) in the determination of which records represent the same vaccination event, as well as how information about a vaccination event is combined from several immunization reports, negatively affects overall data quality and usefulness of registries information. The results of this project are intended to support a uniform alignment of the vaccination level deduplication processes in immunization registries according to recommended guidelines.

For example, it is common for an IIS to receive different and sometimes conflicting information on a vaccination event. Information for a patient on a HIB vaccine and a HIB PRP-T vaccine may be reported on the same day or close to the same day. What are the decision points an IIS needs to know in order to select/consolidate the information for the immunization history of that patient? This document provides a uniform guidance on this topic.

The decisions made during vaccination level deduplication affect the forecasting of vaccine administration according to ACIP recommendations. It is important that input into the vaccine forecasting tool is an accurately recorded immunization history, so that correct and clinically meaningful immunization decisions can be made. Additionally, complete and accurate 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.

Expected benefits

Best practices guidelines on vaccination level deduplication will positively affect immunization registries by providing the following benefits:

- Encourage common deduplication practices, thereby improving overall data quality and usefulness of registry information.
- Serve as a technology-neutral requirement guideline for information technology projects.
- Foster collaboration and aid in communication among IIS professionals.

Intended audience

This document is designed to be read by both programmatic experts and 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 vaccination level deduplication and target actions to address these recommendations.

Intended use

This document contains a set of recommended operational ***best practices*** (including a set of principles to follow and tips to utilize) that are intended for use as a basis for ***requirements*** in vaccination level deduplication algorithms. The specific implementations of deduplication algorithms could vary according to the vendors' technology, application architecture, and specifics of a particular IIS. Additionally, these recommendations can be utilized by IIS for staff training, operational documentation, and communication purposes. Presented approach and results are relevant for and can be used beyond immunization information systems - for developing and documenting best practices and operational requirements for domain-specific deduplication applications in public health, healthcare, and other areas.

Workgroup approach

The workgroup used business engineering and facilitation techniques to analyze immunization registration processes and develop its recommendations. The group used a pragmatic results-oriented business modeling approach that has been effective for modeling of immunization registration and cancer registration operations. Relevant aspects of this approach are outlined below.

- Initial preparatory off-line work (assembling pertinent materials, producing preparatory notes, analysis of processes and development of preliminary drafts) was performed by a small group of business analysts and subject matter experts (SMEs).
- During subsequent face-to-face facilitated modeling session in Washington D.C. (May 8–10, 2006) a full (large) workgroup of SMEs used preparatory materials as a framing/scoping resources and began development and formulation of consensus-based recommendations.
- The post-session work has been aimed at finalizing the development of recommendations. The workgroup was divided into three small groups of SMEs, each addressing a set of remaining tasks during a series of teleconferences. Additional teleconferences were dedicated to reviews of the progress of small groups by the full group of SMEs.
- Facilitation and web-based teleconferencing techniques were used to perform modeling work during the face-to-face workshop and subsequent teleconferences.
- Standard Unified Modeling Language (UML) notation was chosen for this project; this notation is simple enough for the correct intuitive interpretation by public health experts who had not previously worked with UML notation.

- Results of the analysis and the incremental, consensus-based modeling and recommendations development process are captured in the following business modeling artifacts: 1) textual descriptions of principles, rules, restrictions, and operational policies (business rules modeling); 2) process diagrams (UML activity diagrams); 3) diagrams of entities involved in the processes and their relationships (UML domain diagrams); 4) decision tables with deduplication scenarios; 5) other products in tabular and textual formats; and 6) supporting sketches and illustrations.
- The workgroup used the following definition of a consensus among subject matter experts (SMEs) regarding the best practice recommendations developed, which did not reflect 100% agreement, but rather means "*I can live with that and support it.*"

National initiatives of importance to IIS and vaccination level deduplication strategies
The MIROW vaccine deduplication project addresses deduplication issues of vaccination accuracy and credibility of the Immunization Information System (IIS). Public health and health care practitioners rely on IIS systems for timely and correct information on a child's immunization status. In addition, the recommendations may correlate with other national initiatives on broader use of health information technology to affect care delivery and consumer health decisions. Some of these important efforts and how they relate to this deduplication project are summarized below.

I. HHS- AHIC
http://www.hhs.gov/healthit/ahic.html
On September 13, 2005, Health and Human Services (HHS) Secretary Mike Leavitt announced the membership for the American Health Information Community (the Community). The Community is being formed to help advance efforts to reach President Bush's call for most Americans to have electronic health records within ten years. Some of the groups working on various focus areas are:

Electronic Health Record (EHR) Workgroup. Seeks to promote the adoption of electronic health records in physician practices

Personal Health Record Workgroup. Seeks to gain wide-spread adoption of a personal health record that is easy-to-use, portable, longitudinal, affordable, and consumer-centered.

Biosurveillance Workgroup. Will implement the informational tools and business operation to support the real-time nationwide public health event monitoring and rapid response management across public health and care delivery communities and other authorized government agencies.

Immunization Information Systems are relevant to these focus areas of the AHIC in that they are electronic standards-based repositories of clinical and other information regarding immunizations, vaccines and care delivery; they produce a portable personal health record, and they are interoperable and already bridge public health reporting and private practice. The interoperability and reference to exchange standards such as Health Level 7 (HL7) in this document help highlight the importance of standards to connecting to the HHS-AHIC efforts.

II. CCHIT Certification Commission for Healthcare Information Technology
http://www.cchit.org/
CCHIT is the recognized certification authority for electronic health records and their networks, and an independent, voluntary, private-sector initiative. In July 2006 CCHIT announced its first EHR product certification, and additional products will certainly be certified in the future. This initiative is intended to speed adoption of the use of EHRs in small practices by providing a seal of approval to reduce physician risk. Increase in the use of EHR products by providers benefits IIS by producing immunization records of clinical quality and has already been shown to be a source of higher quality records than administrative data.

The vaccination deduplication rules and procedures as defined in this document are designed to be widely adopted by IIS systems and may also have applicability in other EHR systems. These rules and procedures can be used as markers for best practice compliance in both types of systems.

III. PHIN- CDC
http://www.cdc.gov/phin/
The Public Health Information Network (PHIN) is CDC's vision for advancing fully capable and interoperable information systems in the many organizations that participate in public health. PHIN is a national initiative to implement a multi-organizational business and technical architecture for public health information systems. With the acceptance of information technology as a core element of public health, public health professionals are actively seeking essential tools capable of addressing and meeting the needs of the community.

Vaccination level deduplication procedures, business rules and requirements may be areas in which documents such as this can be recognized by PHIN effort. The methods and processes used in this analysis may be adopted as PHIN guidance for soliciting and modeling operational needs and requirements for public health information systems.

The systematic use of subject matter experts, facilitation and business modeling techniques promotes best practices, operational knowledge, and consistent use of quality data across the variety of interrelated public health functional needs. Utilization of business modeling techniques is an important informatics principle that can and should be applied in a variety of operational and system analysis settings.

How these materials are organized
Developed materials are organized in seven chapters and two appendices.

To describe deduplication problems and developed recommendations we needed a set of common agreed upon terms and definitions (Glossary). Such a ***Glossary*** has been developed during the domain modeling activities; it is placed at the end of this document in the ***Appendix A: Domain model***. We recommend using the Appendix A as a reference source for terms and definitions, as well as for a description of the immunization registration area from the perspective of vaccination level deduplication.

Chapter 1: Executive Summary
Provides a concise overview of workgroup's efforts and developed recommendations; intended for public health executives and leaders, as well as for readers who are looking for brief high-level information on this topic. Major workgroup's accomplishments are presented and expected benefits for immunization information systems are listed.

Chapter 2: Process Overview
Describes a vaccination deduplication activities as a three-phase process that includes a Selection phase (identify and group multiple vaccination records that potentially belong to the same vaccination event), an Evaluation phase (evaluate pairs of potentially duplicate immunization records for match/differ decisions), and a Resolution phase (produce a record to represent the vaccination event - select the best record among duplicate records and merge information from duplicate records in a consolidated record). Structurally, the vaccination deduplication process is relatively simple and straightforward, but the complexity and challenges are present in the area of business rules under which this process is conducted. Accordingly, a conclusion has been made that a focus of the work of defining such a process should be placed on the development of business rules (to support deterministic approaches) and recommendations for weights of importance for vaccination variables (to support probabilistic approaches).

Chapter 3: Selection phase
Presents detailed recommendations for selecting potentially duplicate vaccination records based on the developed set of principles and business rules for analysis of relevant vaccination variables. Discussion of the recommended size for the date window that used to group potentially duplicate records together is supported by actual statistics from an immunization registry. Several scenarios illustrate implementation of recommendations for the Selection phase.

Chapter 4: Evaluation phase
Provides guidance on how to determine if vaccination records match (represent the same vaccination event) or differ (represent different vaccination events). Two approaches for evaluating potentially duplicate records are discussed and developed: the sequential approach that applies a set of decision rules for comparisons of variables in vaccination records under consideration and the weights-based approach that compares multiple fields from two records and then computes an aggregate matching score by combining the results of the individual field comparisons. The sequential approach provides a basis for deterministic implementations of vaccination level deduplication algorithms and the weights-based approach provides a basis for probabilistic implementations. Developed recommendations include a list of relevant variables ranked by the level of importance for evaluation of potentially duplicate records, table of distinctive combinations of these variables, principles and business rules, guidance on assigning weights, and formulas to calculate aggregated scores. The chapter concludes with multiple examples that illustrate implementation of sequential and weights-based approaches. A special MS Excel – based automated tool has been developed and used for the purposes of quickly analyzing multitude variations of weights and their impacts on records evaluation. Similar tool could be utilized by IIS during the system development for tuning weights against a set of test cases.

Chapter 5: Resolution phase
Describes recommendations on resolving duplicate records, focusing on two perspectives: those that can be achieved by a) selecting the best record among duplicate records, and b) merging information from duplicate records in a consolidated record. A two-step process should be utilized in the Resolution phase: during the first step, the best record should be selected from duplicate records; this record will be used for clinical purposes. During the second step, using the best record as a base, a new consolidated record (view) should be created (variable-by-variable) that aggregates all available information from duplicate records; this record (view) will be used for both clinical and public health purposes. Extensive discussion of the level of confidence in vaccination data and its influence on a deduplication decision making process is presented; accordingly, a matrix of confidence levels for vaccination data has been developed based on sources of information and other immunization record's variables. Similar to the evaluation phase of the process (chapter 4), sequential (deterministic, rules-based) and weights-based (probabilistic) approaches for the resolution phase have been developed and illustrated by multiple examples.

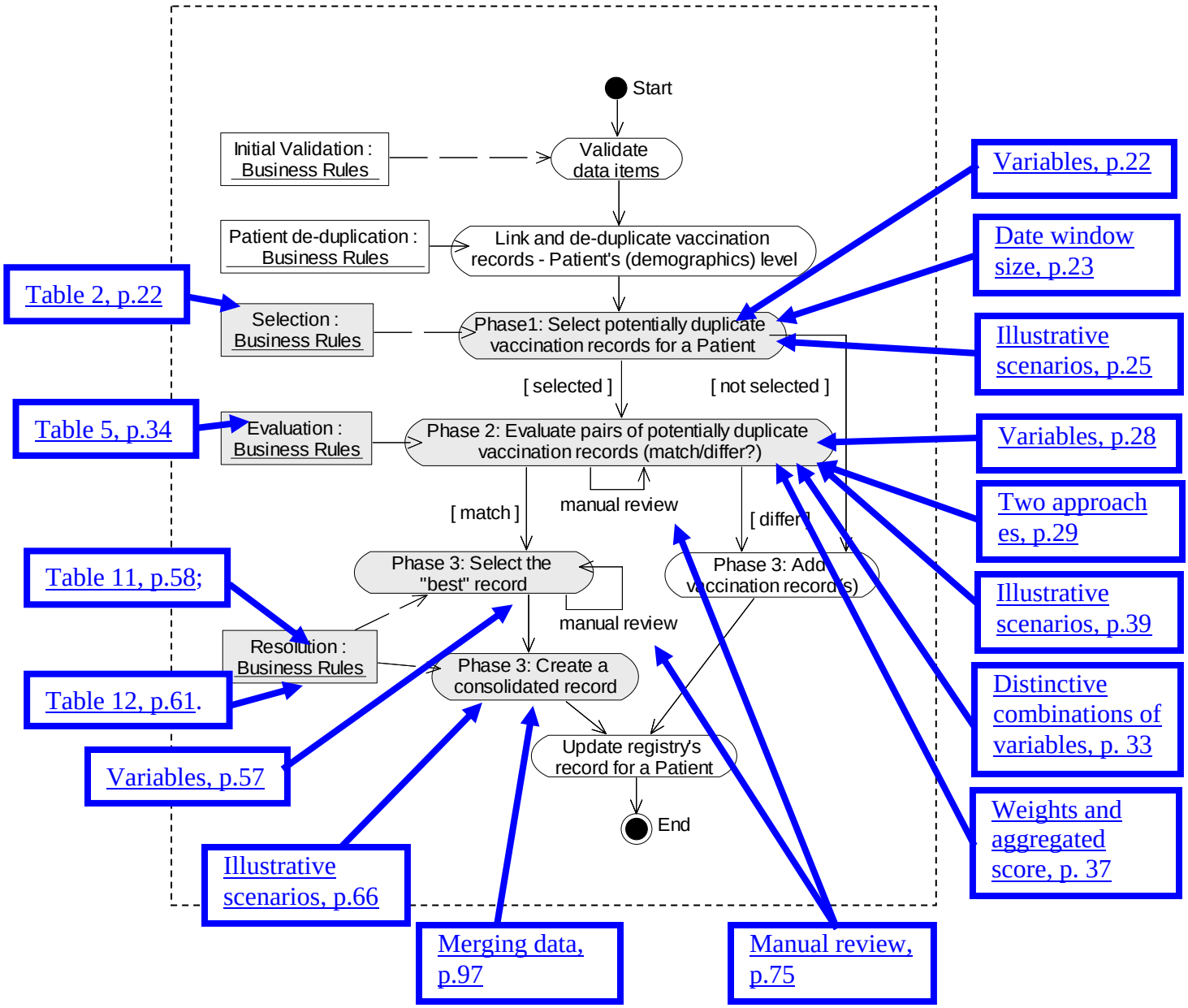
Chapter 6: Additional miscellaneous recommendations
Focuses on the miscellaneous issues relevant to vaccination level deduplication, such as barriers for deduplication and ways to address those barriers, providers-related recommendations, the role of invalid doses (e.g., due to improper storage) with vaccination level deduplication, manual review, and false matches (false positives) and false non-matches (false negatives).

Chapter 7: Conclusions
Reports on the major accomplishments and outcomes of this project. In spite of differences in immunization registry programs, common approaches can be discovered and agreed upon. Business modeling initiative provides an efficient venue for collaboration and exchange of ideas among peers. The results of this project are intended to support a uniform alignment of the vaccination level deduplication processes in immunization registries according to recommended guidelines.

Appendix A: Domain model describes the immunization registration area from the perspective of vaccination level deduplication and provides a set of agreed upon terms and definitions (Glossary). ***Appendix B: Merging data from duplicate records*** presents the issue of creating a consolidated record and relevant considerations in details.

The illustration on the next page places developed recommendations in context of the vaccination level deduplication process and can be used for navigation to major topics and artifacts.

Recommendations in context of the vaccination level deduplication process
(see pp. 20-21 for the process overview diagram and description).



Chapter 1: Executive Summary

The Modeling of Immunization Registry Operations Workgroup (MIROW) was formed by the American Immunization Registry Association (AIRA) in partnership with Centers for Disease Control and Prevention/National Immunization Program (CDC/NIP) to develop a Best Practice guidebook for Immunization Information Systems (IIS).

This document's topic—***vaccination level deduplication***—was selected based on feedback from a survey conducted by the MIROW Steering Committee within the immunization registry/IIS community.

The workgroup, utilizing modern business modeling and facilitation techniques, conducted a 2-and-a-half-day meeting in Washington, D.C. (May 8–10, 2006), as well as multiple pre- and post- meeting teleconferences to analyze existing practices and to formulate consensus-based recommendations. These recommendations are presented as a part of the emerging Best Practice guidebook—a section titled "Vaccination Level Deduplication in Immunization Information Systems".

Major ***accomplishments*** of the workgroup are:

- Defined the vaccination level deduplication process as three-phase process that includes Selection, Evaluation, and Resolution phases.
- Developed principles on which to base the vaccination level deduplication process, business rules to follow, and specific scenarios that illustrate application of principles and business rules.
- Formulated a concept that each vaccination event should be presented in the registry by two records (views) – the best record (view) selected from vaccination reports (utilized for clinical purposes) and the consolidated record (view) that aggregates information from all available sources (utilized for public health and clinical purposes).
- Created specific detailed recommendations for selecting the best record and for constructing the consolidated record.
- Developed a decision matrix for recommendations on levels of confidence based on sources of information.
- Developed and re-confirmed key definitions for vaccination level deduplication.
- Reached consensus among subject matter experts, demonstrating that despite differences in immunization registry programs, common approaches through consensus can be developed and agreed upon with business modeling and facilitation techniques.

Best practices guidelines on vaccination level deduplication will positively affect immunization registries by providing the following ***expected benefits***:

- Encourage common deduplication practices, thereby improving overall data quality and usefulness of registry information.
- Serve as a technology-neutral requirement guideline for information technology projects.
- Foster collaboration and aid in communication among IIS professionals.

The results of this project are intended to support a uniform alignment of the vaccination level deduplication processes in immunization registries according to recommended guidelines.

Recommendations

Chapter 2: Process Overview

A process overview diagram for a vaccination deduplication process is presented in **Figure 1**. Following are the three phases of the process:

Phase 1. SELECTION: Identify and group multiple vaccination records that potentially belong to the same vaccination event.

Phase 2. EVALUATION: Evaluate pairs of potentially duplicate immunization records for match/differ decisions. Possible decisions/outcomes here are:

a) declare the pair a duplicate— both vaccination records represent the same vaccination event;

b) declare the pair not a duplicate —vaccination records represent different vaccination events;

c) declare the pair a possible duplicates (grey zone) requiring manual evaluation.

Phase 3. RESOLUTION: Produce a record to represent the vaccination event—select the best record among duplicate records and merge information from duplicate records in a consolidated record.

This process overview diagram does not specify either a deterministic or probabilistic nature for the approach, reflecting a view that the ***workgroup should not develop a specific algorithm, but rather create a uniform agreed-upon basis for creation of such algorithms (both deterministic and probabilistic).*** Accordingly, the workgroup developed principles, business rules, scenarios, and other materials (e.g., generic measures of importance—weights—for the immunization variables) that could be used in both cases. The resulting formulation of best practice guideline recommendations could be referenced by immunization programs and software vendors in developing specific vaccination level deduplication algorithms.

The process itself is relatively simple and straightforward. The complexity and challenges are in the area of business rules under which process steps are conducted; in other words, this is a rules-driven process. Accordingly, the work of defining such a process has been focused on the development of deterministic business rules, as well as products to support probabilistic approaches (e.g., weights of importance for the variables).

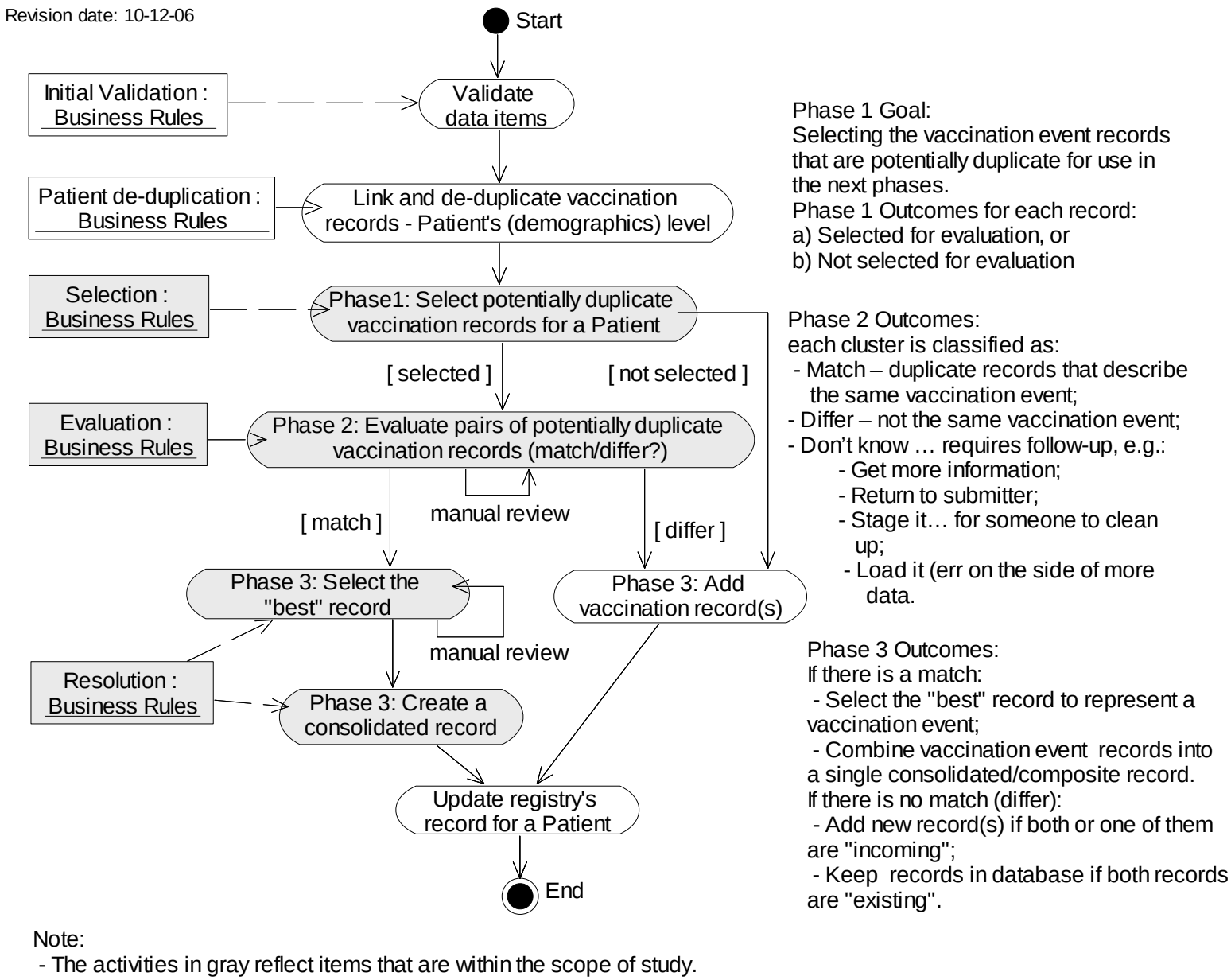


Figure 1. Process overview diagram for the vaccination level deduplication in IIS.

Chapter 3: Selection Phase

The Selection phase is concerned with identifying and grouping records that potentially belong to the same vaccination event. The goal of this phase is to select the potentially duplicate vaccination event records for use in the next phases—Evaluation and Resolution—of the deduplication process.

The expected outcomes of the Selection phase for each examined record are a) Selected for evaluation, or b) Not selected for evaluation. Clusters of records selected at this phase will be examined further in the next phases, Evaluation and Resolution.

Variables used in business rules determining the outcome of the Selection phase are presented in Table 1; principles and business rules for the Selection phase are presented in Table 2. For the detailed description of the variables presented in Table 1, refer to Table A-1 in Appendix A.

Table 1. Variables for the Selection phase.

Variable Name (Entity – Attribute)		Notes
Vaccine	Family/Group Name	The variable "Vaccine – Family/Group Name" is often derived from the "Vaccine - Type" variable. A "Vaccine – Family Group Name" may be matched to more than one "Vaccine – Types." For example, Hib vaccine as Family/Group Name includes the Types Hib-PRP-T, Hib – HbOC, Hib – PRP-OMP, and Hib-Unspecified.
Vaccination Encounter	Date	

Table 2. Principles and business rules for the Selection phase.

P/BR #	Principle / Business Rule Statement	Comments
P04	We would like to be more inclusive than exclusive.	See discussion of false positives versus false negatives on p.76 See discussion regarding the size of the window for selecting candidate duplicate records below.
BR01	If vaccination events for the same Vaccine – Family/Group occur within a maximum window of 23 days, they need to be examined. A registry can set a tighter constraint, based on: • Staffing for manual review;	This business rule is applied first and is the precondition for the use of any other business rules. An explanation for 23 days window: most shots allowed to be made within 28 days of each other, minus 4 days grace period. Not all states use such a grace period (e.g.,

P/BR #	Principle / Business Rule Statement	Comments
	• A trend analysis of the registry data (then it can be constrained appropriately in favor of processing time); • Knowledge of registry's data.	Vermont). Size of the date window should be the same regardless of how data came to the registry (e.g., user interface, electronic upload, etc.), or the source or nature of the data (billing/clinical). Also, see the discussion of the recommended size for the date window – below. Additional conditions should be evaluated to exclude typos-related errors and generic—e.g. month/year, without a date—entries. Examples: illustrative scenarios S001 – S005 at the end of this chapter.
BR02	A record for the vaccination event must be compared with all and any of the vaccination event records with the same Vaccine -Family/Group.	Combination vaccines are related to more than one Vaccine – Family/Group. If there is a partial match within the Vaccine Family/Group Name (a combination vaccine), they need to be examined. Examples: illustrative scenarios S004, S005 at the end of this chapter.
BR03	Identical records should not be selected for deduplication. If there are identical records for the vaccination event, all of them but one has to be deleted.	This business rule addresses the real probability that a provider will continue to send in complete histories for a patient and so the same record will be received many times. Two records can be considered identical if they have identical values for the following minimum set of variables: date, provider, vaccine type. This business rule can not be applied to incomplete records, where some of these data elements are missing.

Discussion of the recommended size for the date window

When selecting groups of potentially duplicate records, two kinds of mistakes can be made. First, records that are truly duplicates can be misidentified as non-duplicates. Second, records that are truly different can be identified as matching. The risk with the first case is that true duplicate records are not evaluated or resolved. The second causes loss of data. (See also the discussion of false-positives vs. false-negatives on p.76).

The decision on a size of the date window to use when selecting groups of immunization records

for further evaluation has a direct impact on the likelihood of making one of the mistakes described above. The narrower window is less likely to select for the evaluation records that are not true duplicates, while a broader window is less likely to miss true duplicates.

According with the business rule BR01, the maximum size of the date window is 23 days. This is a top border; a smaller size window (2–10 days) can be selected by a registry based on the analysis of their data. An example of such an analysis has been provided by CHILD Profile Immunization Registry (Washington). It is based on the examination of state's Medicaid data conducted several years ago. Of those records that appeared to be duplicates but with different dates:

- 55% occurred with a 1 day difference;
- 14% occurred with a 2–5-day difference;
- 11% occurred with a 6–10-day difference;
- 3% occurred with an 11–23-day difference.

Another 11% occurred at exactly 1 month or exactly 2 months later.

The sample size was small, but clearly the bulk – 80% – was within 10 days. These data provide some basis for setting the date window at no more than 10 days.

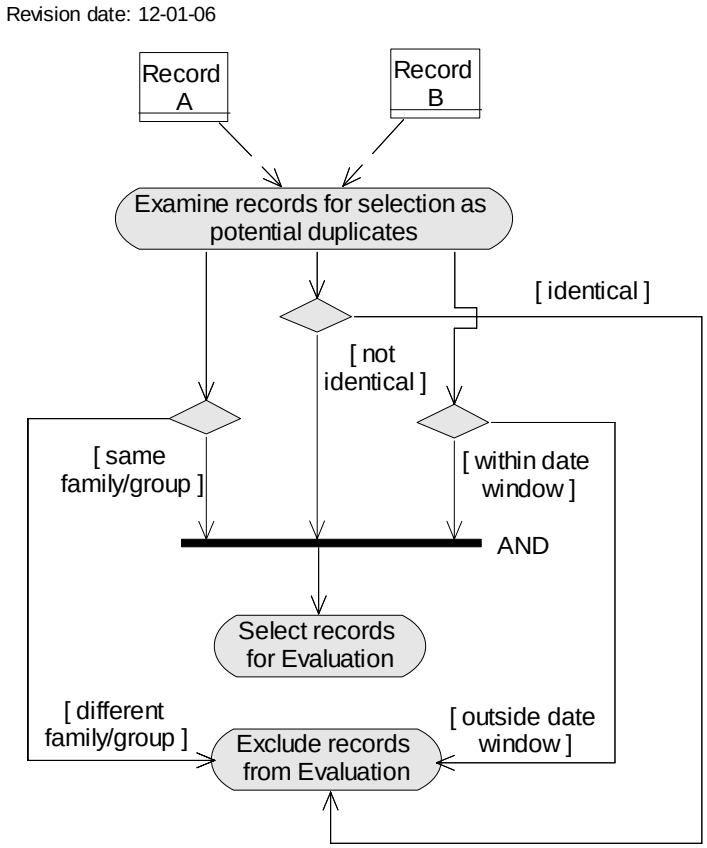


Figure 2. Selection phase overview

Illustrative scenarios

Following are decision tables with some scenarios for the Selection phase of the vaccination level deduplication process that illustrate application of principles and business rules from Table 2.

Scenario ID = S001	**Decision table for the Selection phase**			
	Conditions	The timeframe is within the window		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccine Encounter	Date	05-01-2006	05-24-2006	
Vaccine	Family/Group Name	VAR	VAR	
	Actions	Record A & B are **selected for evaluation**		
	Reasons/Comments	BR01		

Scenario ID = S002	**Decision table for the Selection phase**			
	Conditions	The timeframe exceeds the maximum window of 23 days		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccine Encounter	Date	05-01-2006	05-25-2006	
Vaccine	Family/Group Name	VAR	VAR	
	Actions	Record A & B are **not selected for evaluation**		
	Reasons/Comments	BR01		

Scenario ID = S003	**Decision table for the Selection phase**			
	Conditions	This is an example of two vaccines of different types but same Family/Group		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccine Encounter	Date	05-01-2006	05-01-2006	
Vaccine	Type	HIB-PRP-T (ActHIB®)	HIB – HbOC (HibTITER®)	
Vaccine	Family/Group Name	HIB	HIB	
	Actions	Select for Evaluation		
	Reasons/Comments	BR01 The HIBs match		

Scenario ID = S004	**Decision table for the Selection phase**			
	Conditions	This is the combination example reported as one vaccination event related to two Family/Groups		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccine Encounter	Date	05-01-2006	05-01-2006	
Vaccine	Type	HepB-HIB	HIB	
Vaccine	Family/Group Name	HepB and HIB	HIB	
	Actions	All records with the approximate vaccination encounter date for HepB and HIB are **selected for evaluation**		
	Reasons/Comments	BR01 & BR02 The HIB's match		

Scenario ID = S005	**Decision table for the Selection phase**			
	Conditions	Three (3) vaccination event records; record A is for a combo vaccine (belongs to multiple family/groups). Records B and C each belong to a single family/group that is an individual component of record A's family/groups.		
Entity	**Attribute**	**Record A**	**Record B**	**Record C**
Vaccine Encounter	Date	05-01-2006	05-01-2006	05-01-2006
Vaccine	Family/Group Name	DTaP and Polio and HepB (derived from CPT for Pediarix®)	Polio	HepB
	Actions	All records with the approximate vaccination encounter date for DTaP, HepB and IPV are **selected for evaluation**		
	Reasons/Comments	BR01 & BR02. The IPV and HepB match. The cluster goes forward.		

Chapter 4: Evaluation Phase

The goal of the Evaluation phase of the vaccination level deduplication process is to determine if the pair/cluster of vaccination records ***match*** (represent the same vaccination event) or ***differ*** (represent different vaccination events).

The expected outcome of the Evaluation phase is one of the following conclusions regarding the cluster of potentially duplicate records that have been identified and grouped during the Selection phase:

- Match—duplicate records that describe the same vaccination event
- Differ—not the same vaccination event
- Don't know—possibly requires a manual review (see p. 75) and follow-up to arrive at Match or Differ conclusion, e.g.:
    - Get more information
    - Return to submitter
    - Stage it… for someone to clean up
    - Load it (err on the side of more data) – mark as such

Variables

Variables for making a decision "duplicates/not duplicates" are presented in the **Table 3**; additional variables could be added to this list if necessary. This set of variables is used in determining the outcome of the Evaluation phase**.** For the detailed description of variables included in Table 3, refer to Table A-1 in Appendix A.

Table 3. Variables for the Evaluation phase

Variable Name (Entity – Attribute)		Level of importance for evaluation of potentially duplicate records	Notes
Vaccination Encounter	Date	High (when the same) Low (when different)	
Vaccine	Type	Medium	We also might have CVX code, CPT code.
Vaccine	Trade Name	Medium (when different) Low (when the same)	
Vaccine	Lot Number	High (when different) Low (when the same)	Lot number is present only in 10% of records in Wisconsin, 35% in Michigan, 15% in Minnesota.
Provider Organization	Name	Low	
Vaccination Event Submission	Record Source Type	Medium (when both "administered") Low (when both "historical") Low (when different or absent in one record)	Record Source Type is "Administered" or "Historical" (see item 30 in Table A-1, Appendix A) "Administered" is from Primary Submitter, "Historical" is from Secondary Submitter – see Matrix of confidence levels - Table 9.
Vaccination Event	Compromised Dose (flag)	High	See item 23 in Table A-1, Appendix A

Notes

- Match in lot numbers is not decisive for the records evaluation since the same lot number could be used by a single provider (as well as by different providers) over a period of time.
- "Provider Organization—Name" variable is used throughout this section, but it is not always available (e.g., reports from Medicaid) since the administering provider name is not always included in the data for each vaccination event. On the other hand, the variable "Submitter-Name" is always available and together with the variable "Vaccination Event Submission—Record Source Type" (administered/historical) in many cases could be sufficient to derive the "Provider Organization—Name" variable. See Table A-1 in Appendix A for the variables' description.
- Records selected during the Selection phase belong to the same family/group and have vaccination date within a certain date window.

Two approaches for the Evaluation phase

Sequential (deterministic) approach
The sequential approach (also known as rule-based matching [2]) applies a set of decision rules for comparisons of variables in vaccination records under consideration. These rules evaluate individual variables as well as distinctive combinations of variables. This approach provides a basis for deterministic implementations of vaccination level deduplication algorithms.

A sequential approach can short-circuit the comparison computations by testing most discriminating variables and combinations of variables first. If a rule's condition is true, then a decision match/differ for records is made or additional conditions (rules) are tested. For example, the most discriminating variables—date and lot number—could be analyzed first. If date and lot number are present in both records and are the same for both records, then records can be evaluated as duplicates. More elaborate sequences can be used, but this is just to illustrate an idea of a sequential deterministic approach.

Rules-based and short-circuiting strategies should be supported/justified with the analysis of the IIS-specific data.

Weights-based (probabilistic) approach
The weights-based approach (also known as multi-field matching based on weighted sums [2]) evaluates possible matches by comparing multiple fields from two records and then computing some kind of aggregate matching score by combining the results of the individual field comparisons. Development of this approach provides a basis for probabilistic implementations of vaccination level deduplication algorithms.

For example, for each of the major variables presented in **Table 3** above, a weight can be assigned to indicate how close values of these variables are in two records under examination (e.g., for variable "vaccine lot number": 45 if values are the same, -25 if values are different, for variable "vaccine trade name": 35 if values the same, -5 if values are different, etc). Decision "duplicates/not duplicates" would be made based on the combined quantitative measure of these weights. A proportion between assigned weights is based on the following types of considerations: 1) if both records have lot numbers which differ, the records are less likely to be duplicates; 2) if records are from different providers and each provider claims administering a shot, the records are less likely to be duplicates.

Recommendations presented in this document are at the requirements level. It will be an implementation level task for a software vendor to select weights, based on real data from a specific registry, with utilization of a sound statistical methodology.

Weights assignment strategies should be supported/justified with the analysis of the IIS-specific data.

Discussion: comparison of two approaches
Any deduplication approach, however well thought through, will produce some errors in determining if vaccination records are duplicates or not. Deterministic approaches (such as a

sequential approach described above) try to reduce such errors by adding decision rules that involve analysis of additional variables as well as more sophisticated evaluation logic. Probabilistic approaches (such as a weights-based approach described above) try to reduce such errors by manipulating and fine-tuning weights assigned to outcomes of comparison for individual variables. A good deduplication quality can be achieved with the implementation of each of these approaches.

Both approaches are developed here on a general level; resulting recommendations should be used by IIS and software vendors as requirements for a variety of vaccination level deduplication algorithms that can use a sequential approach, or weights-based approach, or a combination of both (see business rule BR13 in Table 5 for recommendations regarding combining two approaches). **Figure 3** and **Figure 4** illustrates approaches for the Evaluation phase of the process.

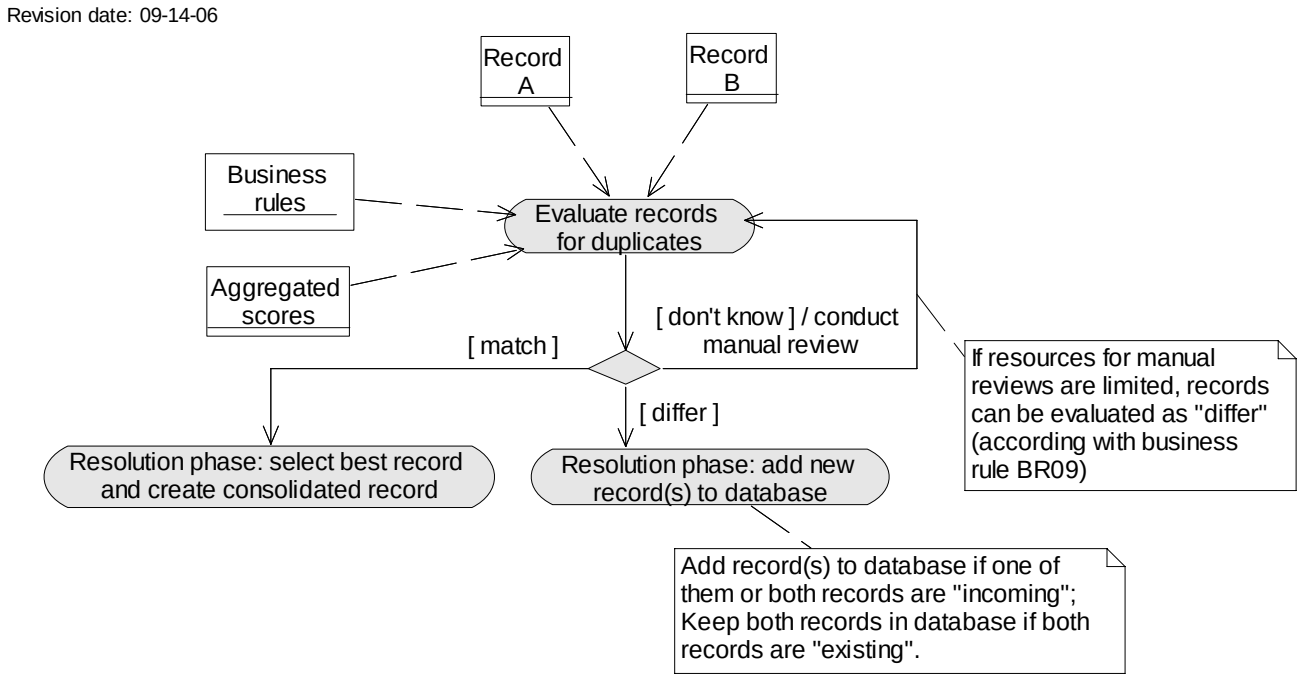


Figure 3. Evaluation phase overview

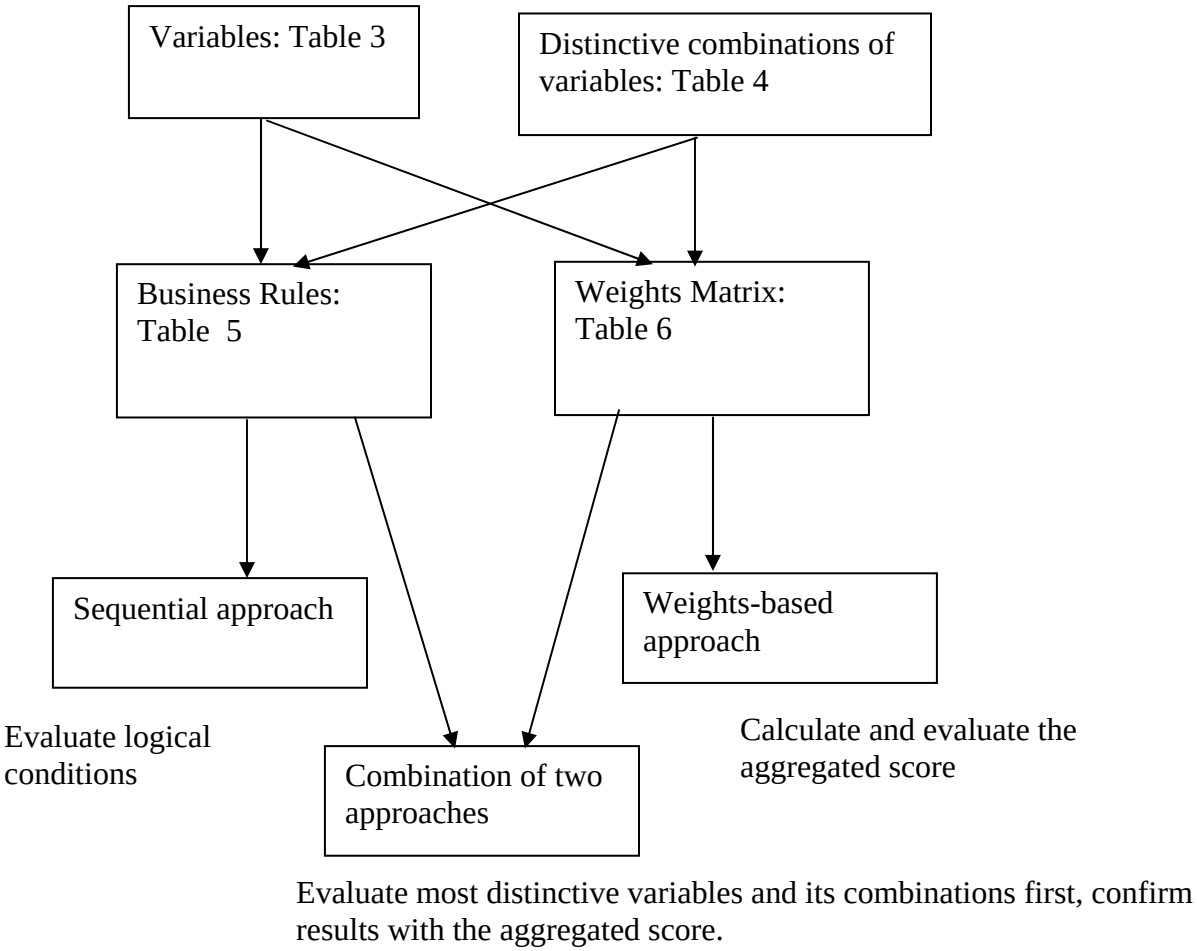


Figure 4. Approaches for the Evaluation phase of the process

Distinctive combinations of variables for the Evaluation phase

Table 3 above provides a general overview of levels of importance for each variable during the Evaluation phase, pointing out the most discriminating variables. The decision making (match/differ for candidates duplicates records) should also take into consideration specific combinations of variables, so most discriminating combinations of variables have to be identified and analyzed. **Table 4** on the next page contains distinctive combinations of variables for the Evaluation phase; it could be used as a basis for both sequential and weights-based approaches.

Table 4. Distinctive combinations of variables for the Evaluation phase.
(S = same, D = different, N – present in one record and absent in another record)

	Date	Lot #	Vac. Type	Trade Name	Provider	Record Source Type	Importance (certainty)	Most probable evaluation outcome	Approx. numerical importance	Comments
1	S	S	S	S	S		High	Match	99	
2	S	D			S		High	Differ	95	
3										
4	D	D					High	Differ	95	Different lot numbers is a high indication of different records, especially when dates are also different.
5	S	N	D		D		High	Match	90	Common to have different vaccine types coming from different sources, such as health plans, historical data.
6	D	S	S	S	S		High	Match	95	

Principles and Business rules

Table 5. Principles and business rules for the Evaluation phase

P/BR #	Principle / Business Rule Statement	Comments
P09	A match in some variables is more important than others.	e.g., lot number vs. vaccine event route: difference in lot numbers is much more important.
P10	The degree of confidence in the data should be taken into consideration.	Additional conditions should be evaluated if confidence in the data quality is low; see matrix of confidence levels, p. 56. See business rule BR15 below.
P11	If vaccination encounter dates are different in records under evaluation, the proximity of these dates has to be taken in consideration.	Records that are 2 days apart are more likely to be duplicates than records that are 22 days apart. A measure of difference should be used for the Vaccine Encounter Date variable since 23-day window is implemented. The bigger the difference between dates, the smaller the weight should be. In other words, for the weights-based approach weights should be adjusted for dates mismatch relative to the degree of difference, e.g. -1 weight if 1 day difference, -10 weight for 23 days difference. Note: Records selected during the Selection phase belong to the same family/group and have vaccination date within a certain date window. See illustrative scenario S016 for the implementation example.
P12	Considerations of front-end vs. back-end processing should not have impact on the decision match/differ for the evaluated records.	The fact that one of the records is incoming and another record is existing, or both records are existing, or both records are incoming has no impact on the decision match/differ for the evaluated records.
P13	Registries should track the variable “Vaccination Event Submission—Record Source Type” (“administered” vs. “historical”) for each record.	Presence of this information could make a decisive difference: see, for example, illustrative scenarios S006 and S006A at the end of this chapter.
BR09	Records selected for evaluation at the Selection phase should be considered different until proven to be duplicates.	At the beginning of evaluation it should be assumed that two records are different; the information that contradicts that assumption should be looked for. Reason: *false-negatives (false non-matches)* are better than *false-positives* (false matches) because false positives result in the loss

P/BR #	Principle / Business Rule Statement	Comments
		of the immunization data. See also the discussion on pp. 76-78. When evaluation of two potentially duplicate records results in the decision "Don't know," manual review should be conducted to clarify the situation in terms Match/Differ. If the manual review can not help to determine the outcome or if registry does not have resources to conduct a manual review, then this business rule BR09 should be applied and these records should be designated as different (not duplicates). This business rule should be applied last, only after evaluation with other business rules is inconclusive. See illustrative scenarios S006, S012 for implementation examples.
BR10	If vaccine lot numbers are different in evaluated records, these records are most likely to be different (not duplicates).	Additional conditions could be evaluated to exclude typos-related errors and generic—e.g. 9999—entries. Lot number provides the most discriminating information, but is present only in ~10% of records in Wisconsin, 35% in Michigan, 15% in Minnesota. The same lot number could be used by a single provider (as well as by different providers) over a period of time; therefore match in lot numbers is not decisive for the records evaluation. At this point, in the Evaluation phase, family/group name is the same since only records that belong to the same family/group were selected during the previous, Selection phase. See illustrative scenarios S007, S008 for implementation examples.
BR11	If vaccination encounter dates are the same in evaluated records, these records are most likely to be duplicates.	Additional conditions could be evaluated to exclude typos-related errors. At this point – in the Evaluation phase - family/group name is the same since only records that belong to the same family/group were selected during the previous – Selection phase. See illustrative scenario S014 for the implementation example.

P/BR #	Principle / Business Rule Statement	Comments
BR12	Distinctive combinations of variables (presented in **Table 4**) should be considered for the evaluation of candidates records.	This business rule is based upon principle P09 and provides general clues for the evaluation of candidate records. See illustrative scenarios S009, S017 for implementation examples.
BR13	High-confidence and/or most discriminating rules (variables and combinations of variables) should be evaluated first. Evaluation sequence: business rules BR10 and BR11 (as well as most distinctive combinations from Table 4) should be applied first; evaluation of variables Vaccine Type, Vaccine Trade Name, and Provider Organization Name could follow.	The following method of combining sequential (deterministic) and weighted (probabilistic) approaches is recommended: evaluate most discriminating variables and combinations of variables first (e.g., lot number, encounter date), then use aggregated score (weights-based) to confirm results of that evaluation.
BR14	Some immunizations are supposed to be given within 2 days of each other.	Examples: rabies, oral typhoid vaccines.
BR15	If Record Source Types are "Administered" in evaluated records and are from different providers, these records are most likely to be different (not duplicates). If Record Source Type is "Administered" in one record and "Historical" in another record and vaccination dates are close (P11), these records are most likely to be duplicates.	See principle P10 above. Record Source Type is "Administered" or "Historical" (see item 30 in Table A-1, Appendix A). "Administered" is from Primary Submitter, "Historical" is from Secondary Submitter – see Matrix of confidence levels - Table 9. If both records are "historical", it does not provide the distinctive information to make a decision match/differ. See illustrative scenarios S0010, S0011, S013, S006A for implementation examples.

Weights and aggregated score

Table 6. Weights for key variables: Evaluation phase

Variable Name (Entity – Attribute)		WEIGHTS		
		Present in both records: SAME	Present in both records: DIFFER	Absent in one or both records
Vaccine	Lot Number	A1	A2	A3
Vaccination Encounter	Date	B1	B2	n/a
Vaccine	Type	C1	C2	C3
Vaccine	Trade Name	D1	D2	D3
Provider Organization	Name	E1	E2	E3
Vaccination Event Submission	Record Source Type	F1 (F1a and F1h)	F2	F3

Notes for assigning weights:

1) General guidance:

- Higher values in individual weights indicate higher probability that records are duplicates; the higher the aggregated score of weights for all variables, the more likely records are duplicates.
- When variable values in records under consideration are the same, in most cases this indicates a stronger support for evaluating records as duplicates.
- When variable values in records under consideration are different, in most cases this indicates stronger support for evaluating records as different (not duplicates) compared with the scenario when data for that variable are missing in one or both records.
  Therefore, B1>B2, C1 > C3 > C2, etc. In some cases (see item 2 below) these considerations do not apply.
- Negative values can be assigned to weights.

2) More specific guidance:

See Table 3 and Table 4 for specific guidance on assigning values for weights. For example, according to Table 3, F1 should have the lowest value when both records are “Administered" (F1a) and higher value then both records are "Historical" (F1h). F2 (one "Administered" and one "Historical”) would have the highest value (most likely - duplicate records). In other words, F2 > F3 >= F1h > F1a ; this is different from all other variables.

3) A measure of difference should be used for the Vaccine Encounter Date variable since 23-day window is implemented. The bigger the difference between dates, the smaller the weight B2 should be. In other words, weight B2 should be adjusted for dates mismatch relative to the

degree of difference, e.g., 65 weight value if 1 day difference, 23 weight value for 23 days difference (see principle P11 and Table 7).

4) The specificity of information for Vaccine Type variable can be taken into consideration: a match between two specific types of vaccine (e.g., both HIB-OMP or both HIB-PRP) should be weighted heavier compared with the match between one specific and one non-specific type (e.g., between HIB-OMP and HIB-NOS – "not otherwise specified"). To achieve that, weight C1 can be manipulated accordingly.

5) See Table 7 and the section "Illustrative scenarios" for numerical examples (summarized in Table 8).

Calculations for the weights-based approach
The weights-based match/differ evaluation of records is conducted according to formulas (1–7) below:

(1) S = A 1(or A2, or A3) + B1 (or B2) + C1 (or C2, or C3) + D1 (or D2, or D3) + E1 (or E2, or E3) + F1(orF2, or F3), where S is aggregated score for a possible match evaluation.

(2) Smax = max(A1, A2, A3) + max(B1, B2) + max(C1, C2, C3) + max(D1, D2, D3) + max(E1, E2, E3) + max(F1, F2, F3), where Smax is maximum aggregated score possible.

(3) Smin = min(A1, A2, A3) + min(B1, B2) + min(C1, C2, C3) + min(D1, D2, D3) + min(E1, E2, E3) + min(F1, F2, F3), where Smin is minimum aggregated score possible.

(4) R = (S – Smin) / (Smax – Smin), where R is relative aggregated score.

(5) If $\mathbf{R > R_H}$, then match—records are duplicates;

(6) If $\mathbf{R < R_L}$, then differ—records are not duplicates;

(7) If $\mathbf{R_L < R < R_H}$, then evaluation is inconclusive (manual review should be conducted or records should be evaluated as not duplicates (see BR09), depending on registry's resources and policies). See the discussion of manual review issues on p. 75.

$\mathbf{R_H}$ **,** $\mathbf{R_L}$ **-** high and low decision thresholds respectively; a statistical analysis of a registry data has to be utilized for a proper selection of values for $\mathbf{R_H}$ and $\mathbf{R_L}$.

Illustrative scenarios

The following scenarios provide some examples of the implementation of sequential (deterministic, rules-based) and weights-based (probabilistic) approaches for the Evaluation phase of the vaccination level deduplication process. These scenarios are not intended to serve as a comprehensive reference source, but rather as illustrative materials for recommendations presented in this section.

Weights used for these scenarios (**Table 7** below) and R_H , R_L values are to a large degree speculative and for illustration purposes only. It will be an implementation level task for a software vendor to select weights and R_H , R_L values, based on real data from a specific registry, with utilization of a sound statistical methodology.

The overview of illustrative scenarios is presented in **Table 8** at the end of this section.

Table 7. Example of weights assignment for key variables: Evaluation phase.

Variable Name (Entity – Attribute)		WEIGHTS		
		Present in both records: SAME	Present in both records: DIFFER	Absent in one or both records
Vaccine	Lot Number	A1=45	A2= -25	A3=25
Vaccination Encounter	Date	B1=80	B2 = 23 (5 days +), 30 (5 days), 38 (4 days), 43 (3 days), 50 (2 days), 65 (1 day)	n/a
Vaccine	Type	C1=50	C2=5	C3=15
Vaccine	Trade Name	D1=35	D2=-5	D3=20
Provider Organization	Name	E1=25	E2=10	E3=15
Vaccination Event Submission	Record Source Type	F1(“administered”) = -7 F1(“historical”) = 15	F2=60	F3=15

Smax = 295; **Smin** = 1; $\mathbf{R_H}$ **=** 0.60, $\mathbf{R_L}$ **=** 0.40.

Note: Weights in Table 7 have been assigned to adhere to general principles discussed in notes for Table 6; specifically, the relations between weights C1 > C3 > C2, F2 > F3 >= F1h > F1a, etc. are maintained.

Presented illustrative scenarios are divided into two sections: first section contains clear-cut scenarios where decisions can be made with a high degree of certainty; the second section contains more complex and challenging scenarios that require additional deliberations.

All scenarios were first resolved with implementation of the rules-based sequential

(deterministic) approach. After that, the weights-based approach was applied; weights were manipulated to achieve the evaluation outcomes that would match the outcomes of the rules-based sequential (deterministic) approach for each illustrative scenario. Implementation examples for the weights-based approach are summarized in **Table 8** at the end of this section. **Table 7** captures the final assignment of weights for all variables.

Table 8 is a snapshot of an automated MS Excel-based tool that has been developed and used for the purposes of quickly analyzing multitude variations of weights and their impacts on records evaluation. Any change in weights' values (top part of the spreadsheet) is automatically propagated to all scenarios (bottom part of the spreadsheet); the outcome for each scenario gets re-calculated and differences between sequential and weights-based approaches' outcomes are highlighted. Similar tools could be utilized by IIS during the system development for tuning weights against a set of test cases; initial values for weights should be assigned as a result of statistical analysis of a particular registry data.

Clear-cut illustrative scenarios

Scenario ID = S 007	**Decision table for the Evaluation phase**			
CONDITIONS		Different dates, same vaccine type, different lot number, same provider organization		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	8/26/2005	9/9/2005	
Vaccine	Family/Group	DTaP	DTaP	
Vaccine	Type	DTaP	DTaP	
Vaccine	CPT Code	90700	90700	
Vaccine	CVX Code	20	20	
Vaccine	Lot Number	C2221AA	C2253AA	
Vaccine	Trade Name			
Provider Organization	Name	Olmsted Med	Olmsted Med	
Vaccination Event Submission	Record Source Type			
EVALUATION CONCLUSION		**Differ**		
REASON/COMMENTS ***Sequential approach:*** BR10- different lot numbers; Different lot numbers is a high indication of different records, especially when dates are also different. ***Weights-based approach:*** S = -25+23+50+20+25+15 = 108; R = (108-1)/(295-1) = 0.364 <0.4; $R<R_L$ → records are not duplicates.				

Aggregated score calculations for the weights-based approach, scenario S007:

Variable	**Status**	**Weight**
Lot Number	differ	-25
Vaccination Encounter Date	differ, 5+ days	23
Vaccine Type	same	50
Vaccine Trade Name	absent	20
Provider Organization Name	same	25
Record Source Type	absent	15
	Aggregated score (sum):	**108**

Scenario ID = S 008	**Decision table for the Evaluation phase**			
CONDITIONS		Different dates, same vaccine type, different lot number, different provider		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	8/26/2005	9/9/2005	
Vaccine	Family/Group	DTaP	DTaP	
Vaccine	Type	DTaP	DTaP	
Vaccine	CPT Code	90700	90700	
Vaccine	CVX Code	20	20	
Vaccine	Lot Number	C2221AA	C2253AA	
Vaccine	Trade Name			
Provider Organization	Name	Olmsted Med	Brainerd Med	
Vaccination Event Submission	Record Source Type			
EVALUATION CONCLUSION		**Differ**		
REASON/COMMENTS ***Sequential approach:*** BR10- different lot numbers; Different lot numbers is a high indication of different records, especially when dates are also different. ***Weights-based approach:*** S = -25+23+50+20+10+15 = 93; R = (93-1)/(295-1) = 0.313 < 0.4; R< R_L → records are not duplicates.				

Aggregated score calculations for the weights-based approach, scenario S008:

Variable	**Status**	**Weight**
Lot Number	differ	-25
Vaccination Encounter Date	differ, 5+ days	23
Vaccine Type	same	50
Vaccine Trade Name	absent	20
Provider Organization Name	differ	10
Record Source Type	absent	15
	Aggregated score (sum):	**93**

Scenario ID = S 009		**Decision table for the Evaluation phase**		
CONDITIONS		Everything the same except date		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	1/6/2006	1/16/2006	
Vaccine	Family/Group	DTaP, HepB, Polio	DTaP, HepB, Polio	
Vaccine	Type	DTaP-HepB-Polio	DTaP-HepB-Polio	
Vaccine	CPT Code	90723	90723	
Vaccine	CVX Code	110	110	
Vaccine	Lot Number	AC21A011CA	AC21A011CA	
Vaccine	Trade Name	Pediarix®	Pediarix®	
Provider Organization	Name	Brainerd Med	Brainerd Med	
Vaccination Event Submission	Record Source Type			
EVALUATION CONCLUSION		**Match**		
REASON/COMMENTS ***Sequential approach:*** BR12: a distinctive combination of variables from Table 4. ***Weights-based approach:*** S = 45+23+50+35+25+15 = 193; R = (193-1)/(295-1) = 0.653 >0.6; R> R_H → records are duplicates. Note: compare with scenario S006 below; it shows how having more information help make a decision.				

Aggregated score calculations for the weights-based approach, scenario S009:

Variable	**Status**	**Weight**
Lot Number	same	45
Vaccination Encounter Date	differ, 5+ days	23
Vaccine Type	same	50
Vaccine Trade Name	same	35
Provider Organization Name	same	25
Record Source Type	absent	15
	Aggregated score (sum):	**193**

Scenario ID = S 010	**Decision table for the Evaluation phase**			
CONDITIONS		Different dates, same vaccine type, same trade name, different providers, both "administered"		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	3/23/2006	3/28/2006	
Vaccine	Family/Group	DTaP, HepB, Polio	DTaP, HepB, Polio	
Vaccine	Type	DTaP-HepB-Polio	DTaP-HepB-Polio	
Vaccine	CPT Code			
Vaccine	CVX Code			
Vaccine	Lot Number			
Vaccine	Trade Name	Pediarix®	Pediarix®	
Provider Organization	Name	Open Cities HC	BabyTracks	
Vaccination Event Submission	Record Source Type	administered	administered	
EVALUATION CONCLUSION		**Differ**		
REASON/COMMENTS ***Sequential approach:*** BR15- both records "administered" and are from different providers: as such they most likely to be different (non-duplicates). ***Weights-based approach:*** S = 25+30+50+35+10-7 =143; R= (143-1)/(295-1) = 0.483; 0.4<0.483<0.6, R_L <R< R_H, and BR09 → records are not duplicates. Note: If one of these records is "historical" and one is "administered" (i.e., one is Primary Submitter and one is Secondary Submitter), then it would be a Match.				

Aggregated score calculations for the weights-based approach, scenario S010:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	differ, 5 days	30
Vaccine Type	same	50
Vaccine Trade Name	same	35
Provider Organization Name	differ	10
Record Source Type	Same, administered	-7
	Aggregated score (sum):	**143**

Scenario ID = S 011	**Decision table for the Evaluation phase**			
CONDITIONS		Different dates, same vaccine type, one has lot number, different providers, one record is "historical", another is "administered"		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	1/2/2006	1/6/2006	
Vaccine	Family/Group	DTaP, HepB, Polio	DTaP, HepB, Polio	
Vaccine	Type	DTap-HepB-Polio	DTap-HepB-Polio	
Vaccine	CPT Code			
Vaccine	CVX Code			
Vaccine	Lot Number		Ac21B037CA	
Vaccine	Trade Name			
Provider Organization	Name	SMDC	Duluth Clinic	
Vaccination Event Submission	Record Source Type	historical	administered	
EVALUATION CONCLUSION		**Match**		

REASON/COMMENTS
Sequential approach: BR15- one record is "historical", another one is "administered"; as such they most likely to be duplicates.
Weights-based approach: S = 25+38+50+20+10+60 = 203; R = (203-1)/(295-1) =0.687 >0.6; R> R_H → records are duplicates.

Note: if SMDC would report this as an "administered" vaccine, then it would be called Different (BR15).

Aggregated score calculations for the weights-based approach, scenario S011:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	differ, 4 days	38
Vaccine Type	same	50
Vaccine Trade Name	absent	20
Provider Organization Name	differ	10
Record Source Type	differ	60
	Aggregated score (sum):	**203**

Scenario ID = S 013	**Decision table for the Evaluation phase**			
CONDITIONS		Different dates, same vaccine type, different providers, one record is "administered", another record is "historical" *Record A—incoming, Record B—existing; both records are from the automated load.*		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	01/10/2005	01/12/2005	
Vaccine	Family/Group	Hep B	Hep B	
Vaccine	Type	Hep B	Hep B	
Vaccine	CPT Code		90748	(HIB/Hep B)
Vaccine	CVX Code	51		(HIB/Hep B)
Vaccine	Lot Number			
Vaccine	Trade Name			
Provider Organization	Name	123	54 (Medicaid)	
Vaccination Event Submission	Record Source Type	administered	historical	
EVALUATION CONCLUSION		**Match**		
REASON/COMMENTS ***Sequential approach:*** BR15 - Record Source Type is "administered" in one record and "historical" in another record, plus vaccination dates are close (P11). ***Weights-based approach:*** S = 25+50+50+20+10+60 = 215; R = (215-1)/(295-1) = 0.728>0.6; R> R_H → records are duplicates.				

Aggregated score calculations for the weights-based approach, scenario S013:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	differ, 2 days	50
Vaccine Type	same	50
Vaccine Trade Name	absent	20
Provider Organization Name	differ	10
Record Source Type	differ	60
	Aggregated score (sum):	**215**

Scenario ID = S 017	**Decision table for the Evaluation phase**			
CONDITIONS		Same dates, different vaccine type, different lot numbers, same provider organization.		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	6/8/06	6/8/06	
Vaccine	Family/Group	DTaP, HepB, Polio	Polio	
Vaccine	Type	DTaP-HepB-IPV	IPV	
Vaccine	CPT Code			
Vaccine	CVX Code	110	10	
Vaccine	Lot Number	C2221AA	C2253AA	
Vaccine	Trade Name	Pediarix®	IPOL®	
Provider Organization	Name	Kidhealth VT	Kidhealth VT	
Vaccination Event Submission	Record Source Type			
EVALUATION CONCLUSION		**Differ**		

REASON/COMMENTS

Sequential approach: BR12 - distinctive combination of variables (row 2, Table 4).

Weights-based approach: S = -25+80+5-5+25+15 = 95; R = (95-1)/(295-1) =0.320<0.4; $R < R_L$ → records are not duplicates.

This example addresses the following situation: some combination vaccines (such as Pediarix® and IPOL®) can be administered on the same day; they should not be considered duplicates.

Aggregated score calculations for the weights-based approach, scenario S017:

Variable	**Status**	**Weight**
Lot Number	differ	-25
Vaccination Encounter Date	same	80
Vaccine Type	differ	5
Vaccine Trade Name	differ	-5
Provider Organization Name	same	25
Record Source Type	absent	15
	Aggregated score (sum):	**95**

Complex illustrative scenarios

Scenario ID = S006	**Decision table for the Evaluation phase**			
CONDITIONS		Complete record vs. record with incomplete information		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	09/29/2004	09/26/2004	
Vaccine	Family/Group	HepB	HepB	
Vaccine	Type	HepB		
Vaccine	CPT Code	90744		
Vaccine	CVX Code	08		
Vaccine	Lot Number	ENG123		
Vaccine	Trade Name	Engerix-B®-Peds		
Vaccination Event Submission	Record Source Type			
EVALUATION CONCLUSION		**Don't know or Differ**		

REASON/COMMENTS

Sequential approach: "Don't know" evaluation outcome could result in a) manual review for registries that have resources available, or b) implementation of rule BR09 ("Differ" since there is no information to prove that these records are duplicates) for registries that don't have resources available for a manual review.

BR09 – records are not proven to be duplicates. BR09 – records are not proven to be duplicates.

Weights-based approach: S = 25+43+15+20+15+15 = 133; R = (133-1)/(295-1) =0.449; 0.4<0.449<0.6, R_L<R< R_H, and BR09 → records are not duplicates.

Mote: This example represents a registry that does not track "administered" vs. "historical" information for immunization records; see principle P13 in Table 5.

First record is complete with immunization information (codes, lot number); it could be assumed that it is likely "administered". Second record is very incomplete; it could be assumed that it is likely "historical". Then, as per BR15, it would be a match. But BR09 outweighs here. Scenario S006A below explores and represents this logic explicitly.

Aggregated score calculations for the weights-based approach, scenario S006:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	differ, 3 days	43
Vaccine Type	absent	15
Vaccine Trade Name	absent	20
Provider Organization Name	absent	15
Record Source Type	absent	15
	Aggregated score (sum):	**133**

Scenario ID = S006A	**Decision table for the Evaluation phase**			
CONDITIONS		Complete record vs. record with incomplete information		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	09/29/2004	09/26/2004	
Vaccine	Family/Group	HepB	HepB	
Vaccine	Type	HepB		
Vaccine	CPT Code	90744		
Vaccine	CVX Code	08		
Vaccine	Lot Number	ENG123		
Vaccine	Trade Name	Engerix-B®-Peds		
Vaccination Event Submission	Record Source Type	administered	historical	
EVALUATION CONCLUSION		**Match**		

REASON/COMMENTS

Sequential approach: BR15 - Record Source Type is "administered" in one record and "historical" in another record; also, vaccination dates are close (P11).

Weights-based approach: S = 25+43+15+20+15+60 = 178; R = (178-1)/(295-1) = 0.602 > 0.6; R> R_H → records are duplicates.

Note: Compare this with scenario S006 above.

Aggregated score calculations for the weights-based approach, scenario S006A:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	differ, 3 days	43
Vaccine Type	absent	15
Vaccine Trade Name	absent	20
Provider Organization Name	absent	15
Record Source Type	differ	60
	Aggregated score (sum):	**178**

Scenario ID = S 012	**Decision table for the Evaluation phase**			
CONDITIONS		Different dates, different vaccine type, one has trade name, different providers.		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	2/13/2006	2/20/2006	
Vaccine	Family/Group	DTaP, HepB, Polio	DTaP	
Vaccine	Type	DTaP-HepB-Polio	DTaP	
Vaccine	CPT Code			
Vaccine	CVX Code			
Vaccine	Lot Number			
Vaccine	Trade Name	Pediarix®		
Provider Organization	Name	Baby Tracks	Aspen Med Grp	
Vaccination Event Submission	Record Source Type			
EVALUATION CONCLUSION		**Don't Know or Differ**		

REASON/COMMENTS

Notes:

1) Client does have corresponding records for polio given on 2/13/2006 and 2/20/2006, but NOT a HepB given on 2/20/2006, only on 2/13/2006

2) This is an example of where it would be helpful to consider the data source. Because the trade name is in Record A, this may be an "administered" vaccine, from a "Primary Submitter". If "Aspen" is submitting Record B as "historical," then it would be a Match (BR15, but dates, on other hand, are not too close). But if they're both being reported as "administered," then records would be called "Different" since they are from different submitters, with equally high level of confidence and the dates are different enough (BR15).

Sequential approach: "Don't know" evaluation outcome could result in a) manual review for registries that have resources for that, or b) implementation of rule BR09 ("Differ" since there is no information to prove that these records are duplicates – for registries that don't have resources for manual review.

Weights-based approach: S = 25+23+5+20+10+15 = 98; R = (98-1)/(295-1) = 0.330 < 0.4; $R < R_L$ → records are not duplicates.

Aggregated score calculations for the weights-based approach, scenario S012:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	differ, 5+ days	23
Vaccine Type	differ	5
Vaccine Trade Name	absent	20
Provider Organization Name	differ	10
Record Source Type	absent	15
	Aggregated score (sum):	**98**

Scenario ID = S 014	**Decision table for the Evaluation phase**			
CONDITIONS		Same date, same vaccine type, one record has a lot number, different providers; *Record A – incoming, Record B- existing; Record A – from UI, Record B- from automated load.*		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	02/02/2006	02/02/2006	
Vaccine	Family/Group	HIB	HIB	
Vaccine	Type	HIB PRP-OMP	HIB PRP-OMP	
Vaccine	CPT Code			
Vaccine	CVX Code	49	49	(HIB PRP-OMP)
Vaccine	Lot Number	2345mm		
Vaccine	Trade Name			
Provider Organization	Name	332	354	
Vaccination Event Submission	Record Source Type	administered	administered	
EVALUATION CONCLUSION		**Match**		
REASON/COMMENTS ***Sequential approach:*** BR11 – dates are the same. This outweighs the BR15 (both records "administered" and from different providers); BR11 takes priority according to levels of importance for variables in Table 3: match in dates has high importance, match in record source type has medium importance (when both records "administered"). ***Weights-based approach:*** S = 25+80+50+20+10-7 = 178; R = (178-1)/(295-1) = 0.602 > 0.6; R> R_H → records are duplicates.				

Aggregated score calculations for the weights-based approach, scenario S014:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	same	80
Vaccine Type	same	50
Vaccine Trade Name	absent	20
Provider Organization Name	differ	10
Record Source Type	Same, administered	-7
	Aggregated score (sum):	**178**

Scenario ID = S 016	**Decision table for the Evaluation phase**			
CONDITIONS		Different dates, same vaccine type, different providers; Record A- historical, Record B - unknown; *Record A – incoming, Record B- existing; Record A –from UI , Record B- from automated load.*		
Entity	**Attribute**	**Record A**	**Record B**	
Vaccination Encounter	Date	09/09/2004	09/10/2004	
Vaccine	Family/Group	HIB	HIB	
Vaccine	Type	HIB	HIB PRP-T	
Vaccine	CPT Code			
Vaccine	CVX Code	17	48	17 - (HIB, not otherwise specified) 48 - (HIB PRP-OMP)
Vaccine	Lot Number			
Vaccine	Trade Name			
Provider Organization	Name	89	90	
Vaccination Event Submission	Record Source Type	historical		
EVALUATION CONCLUSION		**Match**		
REASON/COMMENTS ***Sequential approach:*** Principle P11 applies – dates are very close. Also, we know that one of the records is historical (BR15). ***Weights-based approach:*** S = 25+65+50+20+10+15 = 185; R = (185-1)/(295-1) = 0.626 > 0.6; R> R_H → records are duplicates.				

Aggregated score calculations for the weights-based approach, scenario S016:

Variable	**Status**	**Weight**
Lot Number	absent	25
Vaccination Encounter Date	differ, 1 day	65
Vaccine Type	same	50
Vaccine Trade Name	absent	20
Provider Organization Name	differ	10
Record Source Type	absent	15
	Aggregated score (sum):	**185**

Table 8. Weights-based approach example: overview of illustrative scenarios for the Evaluation phase

	Lot number	Date	Type	Trade name	Provider	Rec Source Type	Aggr. score	R	R(h)	R(L)	
Weights	A	B	C	D	E	F			0.6	0.4	
			Weights								
Same = S	45	80	50	35	25	-7	295	**Smax**			
Different = D	-25	23	5	-5	10	60	1	**Smin**			
Nor present = N	25		15	20	15	15	90				
						15					
						G5-->F1Sa, G6-->F2D, G7-->F3N,G8-->F1Sh					
		1 Day --> 65, 2 days --> 50, 3 days --> 43, 4 days -- >38, 5 days -->30									
			Scenarios						Outcome caculated	Outcome (rules-based)	Comments
S007	D	D	S	N	S	N					
	-25	23	50	20	25	15	108	0.364	**differ**	**differ**	R<R(L)
S008	D	D	S	N	D	N					
	-25	23	50	20	10	15	93	0.313	**differ**	**differ**	R<R(L)
S009	S	D	S	S	S	N					
	45	23	50	35	25	15	193	0.653	**match**	**match**	R>R(H)
S010	N	D	S	S	D	Sa					
	25	30	50	35	10	-7	143	0.483	**differ**	**differ**	R(L)<R<R(H), and BR09
S011	N	D	S	N	D	D					
	25	38	50	20	10	60	203	0.687	**match**	**match**	R>R(H)
S013	N	D	S	N	D	D					
	25	50	50	20	10	60	215	0.728	**match**	**match**	R>R(H)
S017	D	S	D	D	S	N					
	-25	80	5	-5	25	15	95	0.320	**differ**	**differ**	R<R(L)
S006	N	D	N	N	N	N					
	25	43	15	20	15	15	133	0.449	**differ**	**differ**	R(L)<R<R(H), and BR09
S006A	N	D	N	N	N	D					
	25	43	15	20	15	60	178	0.602	**match**	**match**	R>R(H)
S012	N	D	D	N	D	N					
	25	23	5	20	10	15	98	0.330	**differ**	**differ**	R<R(L)
S014	N	S	S	N	D	Sa					
	25	80	50	20	10	-7	178	0.602	**match**	**match**	R>R(H)
S016	N	D	S	N	D	N					
	25	65	50	20	10	15	185	0.626	**match**	**match**	R>R(H)

Chapter 5: Resolution phase

The Resolution phase of the vaccination level deduplication process has two possible outcomes:
1) If records are found to be duplicates (meaning they describe the same vaccination event) during the previous Evaluation phase, then a record to represent the vaccination event must be produced.
2) If records are found not be duplicates (meaning they describe different vaccination events) during the previous Evaluation phase, then a new record must be added to the registry.

Selecting the best record and creating a consolidated record

Recommendations below are focused on two perspectives; those that can be achieved by a) ***selecting*** the ***best record*** among duplicate records, and b) ***merging*** information from duplicate records into a ***consolidated record***. On one hand, clinical records in many states cannot be changed by anybody but their owner; on the other hand, records have to be changed to aggregate information (e.g., adverse events, administrative data, inventory data – lot numbers) from multiple duplicate records and other sources. That leads to a strategy of having two primary sources of information for each vaccination event as a way to satisfy both sides of these requirements. The background on this issue including a Pro and Con description section to each approach is presented in Appendix B.

The creation and adoption of a consolidated record may require some significant retooling of current operations and technical capacity of the registry, but ultimately the creation of a consolidated record at the vaccination level is the only way to ensure that the immunization event is documented comprehensively into a single, accurate immunization record. Initial quantification of the issue has been conducted in one state registry; it was found that approximately 9.4% of the duplicate DTaPs and approximately 20.2% of the duplicate HepBs would have benefited from merging of information from various sources. (Further details are explained in the Appendix B). Therefore, adopting the strategy to create a consolidated record will more precisely address the intent of the NVAC functional standard #12, which as stated is to *promote accuracy and completeness* of registry data by using a data quality protocol to combine all available information into a single, accurate immunization record.

According to the considerations presented above, a ***two-step process*** should be utilized during the Resolution phase:

- In the first step the best record should be selected from duplicate records. This record will be used for clinical purposes.
- In the second step, using the best record as a base, a new consolidated record (view) should be created (variable-by-variable) that aggregates all available information from duplicate records. This record (view) will be used for both clinical and public health purposes.

A provider enrolled in the registry should be able to view and utilize both records—the best record (reported “as-is”) and the consolidated record, that contains additional information from other sources.

The major difference between these two steps is that during the first step variables are analyzed with the aim of selecting the best ***record***, but during the second step the analysis of variables aims to select the best ***value for each variable***.

The diagram on **Figure 5** below illustrates a two-step approach for the Resolution phase.

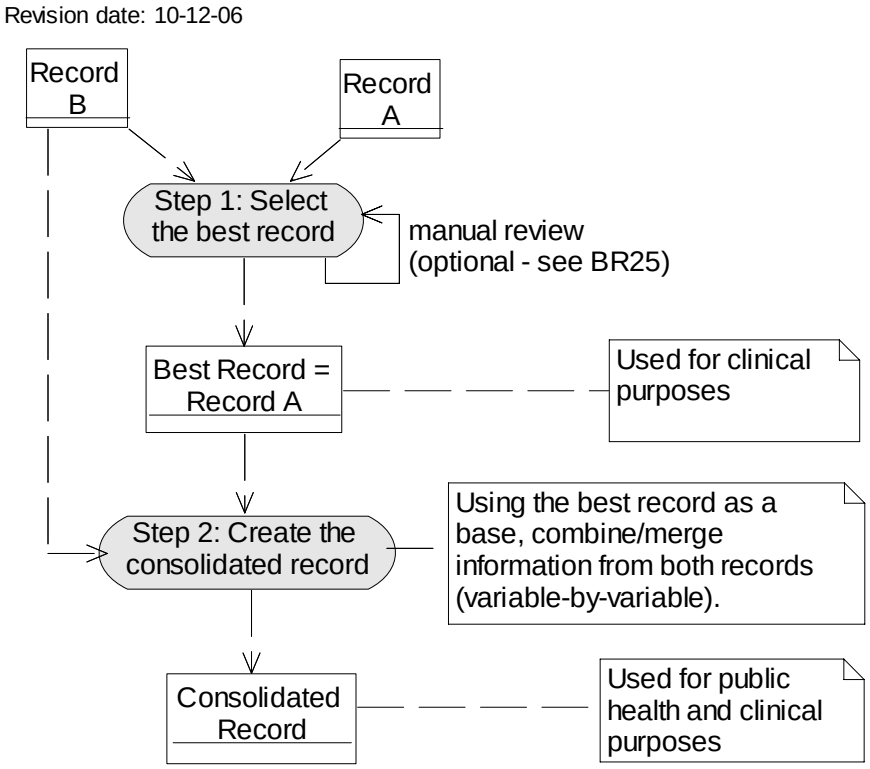


Figure 5. Resolution phase overview —a two-step approach

Discussion of level of confidence in vaccination data
During the Resolution phase records are characterized by the level of confidence or trust regarding their data quality, depending of how these records were submitted to a registry, what type of data records represent, and who submitted them. The confidence level matrix (**Table 9** below) is used for decision making during the Resolution phase of the process for the selection of the best record to represent a vaccination event.

Table 9. Matrix of confidence levels for vaccination data:
based on the attributes of "Vaccination Event Submission" entity

	Method ("How")		Documentation Type ("What")			Record Source Type ("Who")		
Submitters ("Who")	**Electronic Interface (HL7 or other)**	**Registry Specific User Interface**	**Clinical**	**Billing / Claims**	**Transcribed**	**Historical**	**Administered**	**Confidence Level**
Primary Submitter		X	X				X	H+ (high plus)
Primary Submitter	X		X				X	H (high)
Primary Submitter	X			X			X	H - (high minus)
Secondary Submitter	N/A					X		from M to L

Notes:

- Method, Documentation Type and Record Source Type are attributes of "Vaccination Event Submission" entity (see Table A-1 in Appendix A).

- By definition, for Primary Submitter, Record Source Type must be "Administered," and for Secondary Submitter, Record Source Type must be "Historical." Therefore, "Record Source Type" attribute does not play a direct role in determining the confidence level: a Primary/Secondary Submitter attribute already accumulates that "Administered/Historical" information.

- The hierarchy of secondary submitters has to be defined at the local level. For example, in some places nurses work in schools, which raises the confidence level for schools; in other places, where nurses are not present, the confidence level in schools' data is very low.

- If a health care plan reports administered shots for its providers (claims shots as "administered"), then the health care plan is a Primary Submitter (see line 3 of the table above). If a health care plan reports historical data (no claims that "I've administered the shot"), then such a health care plan is a Secondary Submitter (a bottom line of the table above). More detailed explanation follows:

Primary submitters with high confidence levels are providers who submit EMR data or direct data entry into an IIS. Some health plans or health systems transfer data for providers within their organization. In this scenario, the immunization providers utilize the same clinical system and the transfer of data becomes an IT function. These data would have a high confidence level because they are clinical data from a trusted source.

Some health plans submit claims data to an IIS for their participating providers. This would also be considered a primary source but with a lower confidence interval. The health plan sends claims data to the IIS, which is lower quality compared to EMRs or direct data entry.

- Additionally, confidence level can be adjusted based on a personal profile of each provider/submitter (see notes for Table 13 below and the "Providers-related recommendations" discussion on pp. 73-74).

Variables

Variables for selecting the best record to represent a vaccination event are presented in **Table 10**; additional variables could be added to this list if necessary. Resulted set of variables is intended to be utilized in business rules for the Resolution phase**.** For the detailed description of the variables presented in Table 10 refer to the Table A-1 in the Appendix A.

Table10. Variables for the Resolution phase (selecting the best record).

Variable Name (Entity – Attribute)		Level of Importance for selecting the best record	Notes
Vaccine	Type	Low	We also might have CVX code, CPT code.
Vaccine	Trade Name	Low	
Vaccine	Lot Number	Medium	Lot number is present only in 10% of records in Wisconsin, 35% in Michigan, 15% in Minnesota.
Derived variables			
Confidence level		High	See Table 9 for the matrix of confidence levels for vaccination data.
Combo vaccine		Medium	

Sequential (deterministic) approach

Similarly to the Evaluation phase, a ***sequential*** (deterministic) approach can be utilized for the Resolution phase when a set of principles and business rules is applied to 1) determine the best record and 2) create a consolidated record.

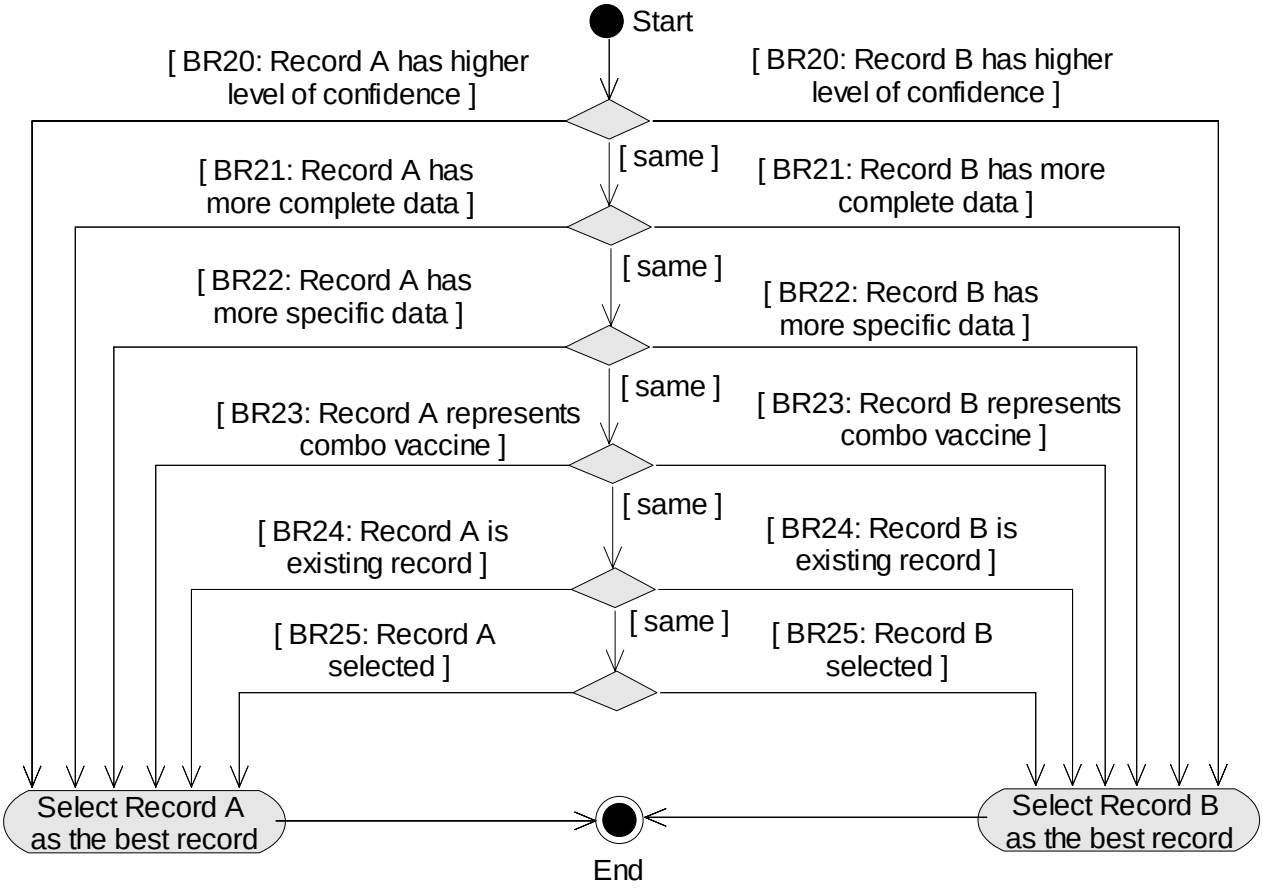
Table 11. Principles and business rules for the Resolution phase (Step 1: selecting the best record)

P/BR #	Principle / Business Rule Statement	Comments
P10	The degree of confidence in the data should be taken into consideration.	This principle has been used in the Evaluation phase as well.
P15	Business rules should be applied completely, in a specified sequence.	The sequence is: BR20→BR21→ BR22→BR23→BR24. See **Figure 6** below for the illustration of sequential implementation of business rules.
BR20	The record with the highest level of confidence should be selected.	See **Table 9** for the matrix of confidence levels for vaccination data. Additionally, confidence level can be adjusted based on a personal profile of each provider/submitter (see notes for Table 13 below and the "Providers-related recommendations" discussion on pp. 73-74). See illustrative scenarios S011-RES, S013-RES, S006A-RES for the implementation examples.
BR21	The record with more complete data should be selected.	The record with more important variables (Table 10) present should be selected. This rule specifies preferences of one variable over another variable: e.g., presence of the lot number is more important than trade name or vaccine type as per Table 10. See illustrative scenario S014-RES for the implementation example.
BR22	The record with more specific data should be selected.	For example, a record with more specific vaccine type—Hib-PRP-T—should be selected over the record with more generic vaccine type—Hib-unspecified. See illustrative scenario S016-RES for the implementation example.

P/BR #	Principle / Business Rule Statement	Comments
BR23	The record that represents a combo vaccine should be selected.	
BR24	The existing record should be selected over the incoming record.	
BR25	Records with earlier or later date should be selected consistently within a particular IIS.	This business rule should be applied only if implementation of business rules BR20-24 has not resulted in selection of the best record. If a particular IIS has resources available, records can be sent to a manual review. If a particular IIS has resources available, selection of the record can be made based on the influence of this selection on the clinical status of the immunization series (see the following notes). Selection of the record with earlier or later date in some cases can affect the clinical status of the vaccine series and lead to the extra-immunization of a patient (extra-immunization is preferred over the under-immunization). <u>Example 1:</u> a child has records of DTP1 shots given on 6/6/05 and on 6/16/05; these records found to be duplicates. DTP2 for that child is given on 7/8/05. If 6/6/05 shot is selected, then DTP2 is valid, but if 6/16/05 shot is selected, then DTP2 is not invalid (22 days between DTP 1-2, the minimum interval of 24 days has been violated). So, the DTP2 shot will have to be repeated in a later case. <u>Example 2:</u> a symmetrical to Example 1 case in which selecting the later date would not lead to a repeated immunization of the patient. One DTP1 dose given on 6/16/05. Then you have DTP2 doses given on 7/8/05 and 7/18/05; and these are found to be duplicates. Choosing the later DTP2 on 7/18/05 means the doses are far enough apart and valid. Choosing the earlier DTP2 would make it invalid and result in it having to be repeated. Here, the earlier choice leads to a repeated immunization, as contrasted with the Example 1.

P/BR #	Principle / Business Rule Statement	Comments
		See illustrative S009-RES for the implementation example. Note: at this stage records under consideration can not be absolutely identical, otherwise they would not be selected as potential duplicates at the beginning of the process during the Selection phase (see BR03 “Identical records should not be selected for deduplication”).

Revision date 12-01-06



Note: This diagram illustrates implementation of a business rules sequence when selecting the best record. “Select Record A (B)” nodes represent possible outcomes and in any given instance only one record (A or B) is actually selected. See the section “Illustrative scenarios” at the end of this chapter for examples.

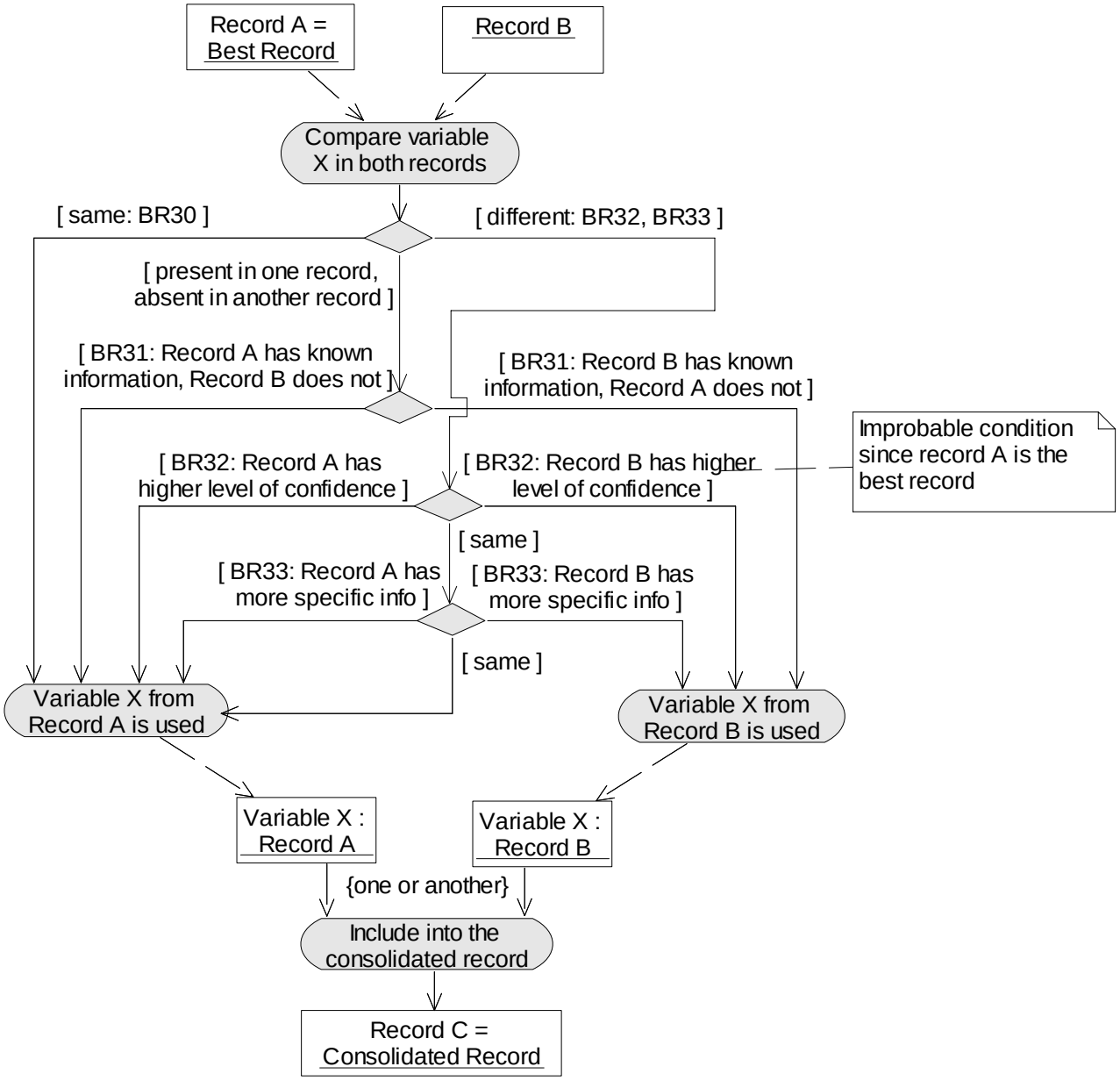
Figure 6. Resolution phase: sequential implementation of business rules to select the best record

Table 12. Principles and business rules for the Resolution phase (Step 2: creating a consolidated record)

P/BR #	Principle / Business Rule Statement	Comments
P18	A consolidated record at the vaccination level that merges all available information from duplicate records and other sources should be created.	Creation of a consolidated record at the vaccination level that merges all available information from duplicate records and other sources is the only way to ensure that the immunization event is documented most comprehensively into a single, accurate immunization record.
BR30	If both records have the same information for a variable, then that information should be used in the consolidated record.	Although seemingly obvious, the rule should be stated explicitly to cover all situations.
BR31	Known information for a variable should be used instead of unknown.	If the best record does not have information for a certain variable and another duplicate record does have that information, the known information for a variable should be used in the consolidated record.
BR32	If duplicate records have different information for a variable, then information from the record with higher level of confidence in data should be incorporated into the consolidated record.	See **Table 9** for the matrix of confidence levels for vaccination data. BR32 has to be applied before BR33.
BR33	If duplicate records have different information for a variable, the more specific information should be incorporated into the consolidated record.	For example, a more specific vaccine type—Hib-PRP-T—should be selected over the more generic vaccine type —Hib-unspecified. Another example could be unspecific date (month and year only) versus a specific (complete) date. See scenario S016-RES for the implementation example.

Note: Business rules presented in Table 12 above address the issue of selecting the best value for each variable in the immunization record.

Revision date 12-01-06



Note: See the section “Illustrative scenarios” at the end of this chapter for examples.

Figure 7. Resolution phase: implementation of business rules to create a consolidated record - merging information for a single variable

Weights-based (probabilistic) approach

This approach allows to weigh the importance of the various variables in duplicate records and compare weighted scores (similar to the weights-based approach in the Evaluation phase). In most cases we are interested in the absence or presence of a data element and not the actual content.

Table 13. Weights for key variables: Resolution phase

Variable Name	Weight	
	Present	Absent
Confidence level: H+ (high plus)	A1	
Confidence level: H (high)	A2	
Confidence level: H - (high minus)	A3	
Confidence level: M (medium)	A4	
Confidence level: L (low)	A5	
Confidence level: unknown		A6
Presence of trade name	B1	B2
Presence of vaccine type	C1 – specific; C3 – non-specific.	C2
Presence of lot number	D1	D2
Is combo vaccine	E1	E2

Notes for assigning weights:

- Higher values in individual weights for a record's variables point to a higher probability that this record will be selected as the best record; the higher an aggregated score of weights for all variables of the record is, the more likely this record will be selected as the best record.
- When the variable is present in the record, its weight should be higher than the weight for the same variable when it is absent from the record. In other word, A1, …, A5 > A6, B1> B2, etc.
- Weights A1, …, A5 are assigned according with the matrix of confidence levels for vaccination data (see Table 9, p. 56). Additionally, personal profile of each provider/submitter can be taken in consideration. A possible way to do it would be a utilization of a profile- related parameter as a multiplier, e.g. Ai adj = Ai * Pj , where Ai - is A1, …, A5, Ai adj - is an adjusted (based on a personal profile for a provider/submitter) value of Ai, Pj - is a profile-related multiplier with values from 0 to 1 (1 – reflects highest level of confidence in that particular provider's data, 0 – lowest level). For example, if A1 = 7 and P = 0,5, then the adjusted value for A1 would be 7 * 0.5 = 3.5. See also "Providers-related recommendations" discussion on pp. 73-74.

Calculations for the weights-based approach
The weights-based selection of the best record is conducted according with formulas (1 - 2) below:

(1) S = A 1(or A2, …, A6) + B1 (or B2) + C1 (or C2) + D1 (or D2) + E1 (or E2),
where S – is aggregated score for a record.

(2A) If **S1 > S2** , or **S1 = S2**, then first record is selected as the best record;
(2B) If **S2 > S1** , then second record is selected as the best record.
(S1 – is aggregated score for the first record and S2 – is aggregated score for the second record).

Table 14. Possible table of weights (used for illustrative scenarios below)

Variable Name	Weight	
	Present	Absent
Confidence level: H+ (high plus)	7	
Confidence level: H (high)	6	
Confidence level: H - (high minus)	5	
Confidence level: M (medium)	4	
Confidence level: L (low)	3	
Confidence level: unknown		3
Presence of trade name	1	0
Presence of vaccine type	3 (specific) 1 (non-specific)	0
Presence of lot number	4	0
Is combo vaccine	1	0

Example:
The first record has Confidence level H+ (clinical, UI, administered), and a lot number and a trade name are present, then the overall score is 12.
The second record has Confidence level H- (billing, electronic interface, administered), and a lot number and a trade name are present, then the overall score is 10.
The first record will be selected as the best record to represent the vaccination event.
See the “Illustrative scenarios” section at the end of this chapter for additional examples.

In many cases a combination of weights-based and sequential approaches could be beneficial for the overall outcome of the deduplication process.

Ownership for the immunization record

Primary Submitter that reported the Vaccination Event owns the record for that vaccination event. (Primary Submitter is a Provider that administered and reported a vaccination – see Table 9 and Table A-1). The ownership means that only this particular Provider can change the record for a vaccination event. Accordingly, administering Provider would have an ownership for the best record and IIS would have an ownership for the consolidated record. Every time the best record is edited or deleted, the deduplication process has to be repeated, so the new best record can be selected and new consolidated record produced.

Some records sent in by Medicaid or health plan may be for providers not yet participating in the IIS and therefore are not owned.

Management of duplicate records

Best practice would be to keep all reported records; some pragmatic strategies have to be developed to address capacities of IIS.

Audit trail should be kept, so all deduplication decisions could be analyzed and, if necessary, reversed.

Illustrative scenarios

Following scenarios provide some examples of the implementation of sequential (rules-based) and weights-based approaches for the Resolution phase of the vaccination level deduplication process. These scenarios are not intended to serve as a comprehensive reference source, but rather as illustrative materials for recommendations presented in this section. Illustrative scenarios presented in this section are continuations of the Evaluation phase scenarios where the outcomes of evaluation were Match.

Scenario ID = S 009 - RES		**Decision table for the Resolution phase**		
CONDITIONS		Everything the same except date		
Entity	**Attribute**	**Record A**	**Record B**	**Consolidated Record**
Vaccination Encounter	Date	1/6/2006	1/16/2006	1/16/2006
Vaccine	Family/Group	DTaP, HepB, Polio	DTaP, HepB, Polio	DTaP, HepB, Polio
Vaccine	Type	DTaP-HepB-Polio	DTaP-HepB-Polio	DTaP-HepB-Polio
Vaccine	CPT Code	90723	90723	90723
Vaccine	CVX Code	110	110	110
Vaccine	Lot Number	AC21A011CA	AC21A011CA	AC21A011CA
Vaccine	Trade Name	Pediarix	Pediarix	Pediarix
Provider Organization	Name	Brainerd Med	Brainerd Med	Brainerd Med
Vaccination Event Submission	Record Source Type			
EVALUATION CONCLUSION		**Match**		
RESOLUTION ACTIONS				
Sequential approach: BR25 - Record B is selected as the best record. A consolidated record (see above) is identical to the Record B - BR30. ***Weights-based approach:*** S1=S2 = 3+1+3+4+1 =12. S1=S2 → Record B is selected as the best record. Note: as per BR25 comments - if a particular IIS has resources available, cases that affect the clinical status of the vaccination series can be sent to a manual review.				

Aggregated score calculations for the weights-based approach, scenario S009 - RES:

Variable/Condition	**Record A**		**Record B**	
	Status	**Weight**	**Status**	**Weight**
Confidence Level	unknown	3	unknown	3
Presence of trade name	present	1	present	1
Presence of vaccine type	present, specific	3	present, specific	3
Presence of Lot number	present	4	present	4
Is combo vaccine?	yes	1	yes	1
Aggregated score (sum):		**12**		**12**

Scenario ID = S 011 - RES		**Decision table for the Resolution phase**		
CONDITIONS		Different dates, same vaccine type, one has lot number, different providers, one record is "historical", another – is "administered".		
Entity	**Attribute**	**Record A**	**Record B**	**Consolidated Record**
Vaccination Encounter	Date	1/2/2006	1/6/2006	1/6/2006
Vaccine	Family/Group	DTaP, HepB, Polio	DTaP, HepB, Polio	DTaP, HepB, Polio
Vaccine	Type	DTap-HepB-Polio	DTap-HepB-Polio	DTap-HepB-Polio
Vaccine	CPT Code			
Vaccine	CVX Code			
Vaccine	Lot Number		Ac21B037CA	Ac21B037CA
Vaccine	Trade Name			
Provider Organization	Name	SMDC	Duluth Clinic	Duluth Clinic
Vaccination Event Submission	Record Source Type	historical	administered	administered
EVALUATION CONCLUSION		**Match**		
RESOLUTION ACTIONS				
Sequential approach: BR20 - Record B is selected as the best record (higher level of confidence in data). A consolidated record (see above) is identical to the Record B – BR30, BR31, BR32. ***Weights-based approach:*** S1 = 4+0+3+0+1 = 8; S2 = 6+0+3+4+1 = 14. S2>S1 → Record B is selected as the best record.				

Aggregated score calculations for the weights-based approach, scenario S011 - RES:

Variable/Condition	**Record A**		**Record B**	
	Status	**Weight**	**Status**	**Weight**
Confidence Level	medium	4	high	6
Presence of trade name	absent	0	absent	0
Presence of vaccine type	present, specific	3	present, specific	3
Presence of Lot number	absent	0	present	4
Is combo vaccine?	yes	1	yes	1
Aggregated score (sum):		**8**		**14**

Scenario ID = S 013 - RES		**Decision table for the Resolution phase**		
CONDITIONS		Different dates, same vaccine type, different providers, one record is "administered", another record is "historical"; *Record A – incoming, Record B- existing; both records are from the automated load.*		
Entity	**Attribute**	**Record A**	**Record B**	**Consolidated Record**
Vaccination Encounter	Date	01/10/2005	01/12/2005	01/10/2005
Vaccine	Family/Group	Hep B	Hep B	Hep B
Vaccine	Type	Hep B	Hep B	Hep B
Vaccine	CPT Code		90748 (HIB/Hep B)	90748 (HIB/Hep B)
Vaccine	CVX Code	51 (HIB/Hep B)		51 (HIB/Hep B)
Vaccine	Lot Number			
Vaccine	Trade Name			
Provider Organization	Name	123	54 (Medicaid)	123
Vaccination Event Submission	Record Source Type	administered	historical	administered
EVALUATION CONCLUSION		**Match**		
RESOLUTION ACTIONS				
Sequential approach: BR20 - Record A is selected as the best record (higher level of confidence in data). A consolidated record (see above) is identical to the Record A with CPT Code added – BR30, BR31, BR32. ***Weights-based approach:*** S1 = 6+0+3+0+0 = 9; S2 = 4+0+3+0+0 = 7. S1>S2 → Record A is selected as the best record.				

Aggregated score calculations for the weights-based approach, scenario S013 - RES:

Variable/Condition	**Record A**		**Record B**	
	Status	**Weight**	**Status**	**Weight**
Confidence Level	high	6	medium	4
Presence of trade name	absent	0	absent	0
Presence of vaccine type	present, specific	3	present, specific	3
Presence of Lot number	absent	0	absent	0
Is combo vaccine?	no	0	no	0
Aggregated score (sum):		**9**		**7**

Scenario ID = S006A - RES		**Decision table for the Resolution phase**		
CONDITIONS		Complete record vs. record with incomplete information		
Entity	**Attribute**	**Record A**	**Record B**	**Consolidated Record**
Vaccination Encounter	Date	09/29/2004	09/26/2004	09/29/2004
Vaccine	Family/Group	HepB	HepB	HepB
Vaccine	Type	HepB		HepB
Vaccine	CPT Code	90744		90744
Vaccine	CVX Code	08		08
Vaccine	Lot Number	ENG123		ENG123
Vaccine	Trade Name	Engerix-B-Peds		Engerix-B-Peds
Vaccination Event Submission	Record Source Type	administered	historical	administered
EVALUATION CONCLUSION		**Match**		
RESOLUTION ACTION				
Sequential approach: BR20 - Record A is selected as the best record (higher level of confidence in data). A consolidated record (see above) is identical to the Record A – BR30, BR31, BR32. ***Weights-based approach:*** S1 = 6+1+3+4+0 = 14; S2 = 4+0+0+0+0 = 4. S1>S2 → Record A is selected as the best record.				

Aggregated score calculations for the weights-based approach, scenario S006A - RES:

Variable/Condition	**Record A**		**Record B**	
	Status	**Weight**	**Status**	**Weight**
Confidence Level	high	6	medium	4
Presence of trade name	present	1	absent	0
Presence of vaccine type	present, specific	3	absent	0
Presence of Lot number	present	4	absent	0
Is combo vaccine?	no	0	no	0
Aggregated score (sum):		**14**		**4**

Scenario ID = S 014 - RES		**Decision table for the Resolution phase**		
CONDITIONS		Same date, same vaccine type, one record has a lot number, different providers; *Record A – incoming, Record B- existing; Record A – from UI, Record B- from automated load.*		
Entity	**Attribute**	**Record A**	**Record B**	**Consolidated Record**
Vaccination Encounter	Date	02/02/2006	02/02/2006	02/02/2006
Vaccine	Family/Group	HIB	HIB	HIB
Vaccine	Type	HIB PRP-OM	HIB PRP-OM	HIB PRP-OM
Vaccine	CPT Code			
Vaccine	CVX Code	49 (HIB PRP-OMP)	49 (HIB PRP-OMP)	49 (HIB PRP-OMP)
Vaccine	Lot Number	2345mm		2345mm
Vaccine	Trade Name			
Provider Organization	Name	332	354	332
Vaccination Event Submission	Record Source Type	administered	administered	administered
EVALUATION CONCLUSION		**Match**		
RESOLUTION ACTION				
Sequential approach: BR21 - Record A is selected as the best record (more complete data – lot number is present in Record A). A consolidated record (see above) is identical to the Record A – BR30, BR31. ***Weights-based approach:*** S1 = 7+0+3+4+0 = 14; S2 = 6+0+3+0+0 = 9. S1>S2 → Record A is selected as the best record.				

Aggregated score calculations for the weights-based approach, scenario S014 - RES:

Variable/Condition	**Record A**		**Record B**	
	Status	**Weight**	**Status**	**Weight**
Confidence Level	high plus	7	high	6
Presence of trade name	absent	0	absent	0
Presence of vaccine type	present, specific	3	present, specific	3
Presence of Lot number	present	4	absent	0
Is combo vaccine?	no	0	no	0
Aggregated score (sum):		**14**		**9**

Scenario ID = S 016 - RES		**Decision table for the Resolution phase**		
CONDITIONS		Different dates, same vaccine type, different providers; Record A- historical, Record B - unknown; *Record A – incoming, Record B- existing; Record A –from UI , Record B- from automated load.*		
Entity	**Attribute**	**Record A**	**Record B**	**Consolidated Record**
Vaccination Encounter	Date	09/09/2004	09/10/2004	09/10/2004
Vaccine	Family/Group	HIB	HIB	HIB
Vaccine	Type	HIB	HIB PRP-T	HIB PRP-T
Vaccine	CPT Code			
Vaccine	CVX Code	17 (HIB, not otherwise specified)	48 (HIB PRP-OMP)	48 (HIB PRP-OMP)
Vaccine	Lot Number			
Vaccine	Trade Name			
Provider Organization	Name	89	90	90
Vaccination Event Submission	Record Source Type	historical		
EVALUATION CONCLUSION		**Match**		
RESOLUTION ACTION				
Sequential approach: BR22 - Record B is selected as the best record (more specific data – vaccine type - in Record B). A consolidated record (see above) is identical to the Record B – BR30, BR33. ***Weights-based approach:*** S1 = 4+0+1+0+0 = 5; S2 = 3+0+3+0+0 = 6. S2>S1 → Record B is selected as the best record.				

Aggregated score calculations for the weights-based approach, scenario S016 - RES:

Variable/Condition	**Record A**		**Record B**	
	Status	**Weight**	**Status**	**Weight**
Confidence Level	medium	4	unknown	3
Presence of trade name	absent	0	absent	0
Presence of vaccine type	present, non-specific	1	present, specific	3
Presence of Lot number	absent	0	absent	0
Is combo vaccine?	no	0	no	0
Aggregated score (sum):		**5**		**6**

Chapter 6: Additional Miscellaneous Recommendations

Barriers for deduplication and ways to address those barriers

Policy

- Vaccination record deduplication may not be a high priority project for registries. Usually registries have an algorithm that validates doses and drops duplicate (invalid) doses. In most cases duplicate vaccinations do not particularly affect the child's record in a clinical sense, so often this project is not seen as high priority, and duplicates are left in the record.

- Although it is relatively easy to come up with a set of rules to suppress/delete immunization records that appear to be duplicates, registries feel they have a responsibility to their data contributors, and perhaps are reluctant to delete data. Is the registry closer (more similar) to a medical record or to a school record?

- It's easy to be perplexed about which immunization events to delete, especially when one starts thinking of the potential exceptions to rules- [what if it **was** an OPV and the child was born somewhere else, what if it was a subpotent shot, and was given again, what if the provider didn't know what they were doing, etc.] One has to make some difficult decisions and then live with a certain amount of small error.

Address: Registries need to be aware that no matter how much data quality checking they perform beforehand, sooner or later they will have to think about how to address vaccination level deduplication. Registry staff should decide how they want to handle this problem ahead of time and be able to defend the decision. The decision they make will affect the technical approach.

Technical:

- The registry needs to store the necessary variables to help make deduplication decisions: for instance, store vaccine formulation (Comvax®, Prevnar®, etc., not just general vaccine series like DTaP, pneumococcal, etc.), store the source of the immunization and other relevant variables.

- Some registries choose not to delete shots but to retain everything and just display the best record. In order to do that the registry needs an appropriate built-in capability.

Address: Fortunately, most registries incorporate some or most of vaccination level fields that will help make that decision. Usually the existing fields will suffice to make a certain number of deduplication decisions. If some critical fields are missing, the registry may want to consider adding them to their tables.

Resources:

- Vaccination level deduplication requires technical, programmatic, staff and monetary resources to be implemented and maintained.

- (a) program and technical people to develop the deduplication rules;
- (b) technical people to implement the rules in the registry ;
- (c) staff time to maintain this activity (for example: (i) in order to develop a trusted vs. non-trusted profile, time must be devoted to evaluate the data submitters on an ongoing basis; (ii) human review?
- (d) Money to do all of the above!

Address: If the policy decision is to address this problem then technical, staff, programmatic and monetary resources should be allocated.

Provider-related recommendations

- Include education for providers to reduce the number of duplicate and poorly specified entries registries receive (e.g. improbable vaccinations, wrong billing codes).
- Pre-qualify submitters by closely reviewing initial submissions. For example, qualify submitter based on the quality of their first 3 submissions.
- Registries should establish data quality profiles for each submitter.

Discussion:
Provider communication regarding duplicate vaccine issues is important. Methods like email, newsletters, websites, and the registry application itself can direct and remind providers to enter data on their patients with care to avoid future vaccine deduplication issues. Statements that could be included in such communication include:

- Providers are responsible for the quality of their own data.
- The registry will only work as well as the data that it holds.
- The highest quality data is usually that which is transcribed the fewest times, thus the occurrence of fewer “fat finger” data entry errors.
- Registries do have certain safeguards to prevent duplicate shot entries (for example, messages warning users that an immunization has already been reported for that series within the past 5-day window). Following techniques can be used:
    - Avoid using “unspecified” (a.k.a. NOS) vaccine codes whenever possible. Refer to the list of currently manufactured vaccines and their associated CPT4/Vaccine/Manufacturer codes to ensure accurate reporting
    - Periodically compare patient’s charts to the registry, checking that information (vaccine administration date, vaccine name, lot number, manufacturer, etc.) is entered accurately.
    - Report data from clinical data sources, such as an Electronic Medical Record (EMR). If reports are made to the registry using data extracted from a billing system, audit to ensure that vaccine administration dates and vaccine codes are correct.
- Remember that the immunization data reported are used for various functions by the reporter (IIS) as well as other registry users, including vaccine inventory, and the use of IIS for vaccine ordering and accountability (VODS) population-based immunization profile reports, HEDIS reporting, and school/childcare compliance reporting.

It is advisable to establish a pre-qualification process for those who submit data electronically to the registry. One recommendation was to qualify based on quality of the first 3 submissions.

Here is an example of a profiling process:
1. The submitter will provide a live (not test data) immunization data file for review. The file should contain at least a month's worth of current immunization records.
2. The file will be tested for format and content. Current vaccines and manufacturers should be represented in the file.
3. The registry will establish profiles on the quality of data represented by each submitter. The ratings will be based upon several items, including:

- clinical vs. billing data
- current + historical data, vs. current only
- chart audit comparing data, includes percentage that was transferred on a timely basis
- vaccine inventory information, if applicable

4. The submitter will then be approved for regular electronic transfer to the registry, realizing that their data quality profile rating could affect the priority of vaccine deduplication rules.

Because of changes in personnel that support electronic immunization data reporting systems, there is great potential for the contents and quality of the data to change over time. Periodic review of the electronic reporting systems is necessary to allow for gauging changes to a previous data quality profile rating.

The role of invalid doses (e.g., due to improper storage) in vaccination level deduplication

Doses noted as potentially sub-potent are typically flagged as invalid because the vaccine that was administered has expired or has been stored or administered incorrectly. The Advisory Committee for Immunization Practice (ACIP) states:
Vaccines that have been mishandled (e.g., inactivated vaccines and toxoids that have been exposed to freezing temperatures) or that are beyond their expiration date should not be administered. If mishandled or expired vaccines are administered inadvertently, they should not be counted as valid doses and should be repeated, unless serologic testing indicates a response to the vaccine. (http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5102a1.htm)

An invalid dose due to subpotency and improper immunization practices should not be considered a duplicate shot by a registry's vaccine deduplication algorithm. In accordance with ACIP guidelines, if doses administered are of questionable potency, these doses should not be counted as valid and should be repeated. To avoid confusion, these doses should not be deleted from the IIS, but instead should be flagged in the IIS's database as invalid, suppressed in the forecasting algorithms, and the flag factored into the vaccine deduplication algorithm and process. Doses given as revaccinations will appear as normal doses in the record.

***Recommended process for flagging doses*:**
The flag should be added to the IIS as soon as possible after the Immunization Program learns of the potentially subpotent vaccine. The flag should be placed with the questionable dose in the IIS database and alongside forecasting messages. Additionally, text fields can be added to the database to allow for further explanation of the flag (e.g., "this dose given too soon", etc.). The

clinic administering the original dose will ideally work in full cooperation with IIS staff to flag the appropriate records; if the clinic does not choose to work with IIS staff, clinic staff should be notified that suspect doses have been flagged.

At a minimum, the invalid flag should be added to the database tables associated with vaccine doses administered. However, adding an invalid flag to the vaccine inventory module of the IIS would be most ideal. Lot numbers for vaccines that need to be recalled because of improper storage, manufacturer error, or vaccine administration practices could be flagged as invalid in the vaccine inventory all at once rather than using the more labor intensive process of flagging doses in other tables of the database. The flagging process for invalid doses could also be automated by factoring in the expiration date of a vaccine in the IIS's vaccine inventory module with the date of vaccine administration. For those rare occurrences when doses may have been administered past their expiration date, these doses could be automatically flagged as invalid in the IIS.

Manual review

When doing automatic deduplication of vaccination event records, invariably, there will be cases when it will be impossible for the system to decide whether two records represent the same event. It is possible, for example, that both potentially duplicate immunizations come from un-trusted sources, are equally specific (or nonspecific) and have the same exact information, except for the vaccine dates which are slightly different. When the system is unable to make a decision automatically, those cases are sent to human review.

Manual review: This process involves a member of the registry staff manually examining the records to determine whether they represent the same immunization event. This determination can be made either (a) by the user simply examining the records and identifying an element, so far not incorporated in the automated model, that helps make the determination, and/or (b) by calling the provider who submitted the information to double check its validity, and edit out the immunization that was submitted incorrectly.

Manual review of immunizations is a tedious and time-consuming task. In addition, calling a provider may not necessarily help resolve the duplicate if, for instance, the provider is a secondary source of information. Given that resources are usually limited, the registry has to decide under which circumstances these potential duplicates should be reviewed and resolved by a person. The group recommends the following approaches to reduce the volume of potential duplicate records sent to manual review:

1. Identify the subset of candidate duplicates to send to manual review as being clinically significant. This can be done using the registry's evaluation algorithm. Most registries use algorithms that disregard/invalidate immunizations that are invalid due to minimum interval or age violations (they are noted as not valid and not used to attain completion of a series), and determine whether an immunization group is complete or up-to-date (UTD). ***Using this algorithm***, the registry can:

    a. Review only potentially duplicate immunizations which belong to a series which is not complete (UTD). Since duplicate immunizations are discounted by the

algorithm, if a vaccination series is not UTD, then it's worth examining the series to see if one of the two duplicate immunizations is truly a duplicate and is not a mistyped immunization.

b. Review potentially duplicate immunizations that can affect other doses in the group: for example, if "a" is selected, it would invalidate the next dose, whereas if "b" is selected it would not. In this instance, it is important to know which immunization was really given. This approach would probably involve additional programming on the registry's part. (*Group recognized that this is not an ideal approach, but might be the only reasonable alternative for a large database with a lot of providers - e.g. NYC and Kaiser - where the investment in programming might be worth the reduction in manual work; it's not practical to resolve manually all the shots that the automatic program can't, especially for large registries that process thousands of immunizations a day.*).

2. During manual review, staff may be able to identify patterns of duplicates, perhaps a certain type of duplicate consistently submitted by one specific source. Identifying those "systematic" duplicates can be extremely helpful and the result in an automated resolution that affects a large number of records that can be dealt with at once. (For example, a particular source always sends their immunizations as of the date they were ordered and not the date they were given. The resolution may be that the registry writes a program that automatically identifies and deletes duplicate immunizations from that specific source.) This way a large number of duplicate immunizations may be automatically resolved.

3. If the automated deduplication model used is probabilistic (weights-based approach), review only a narrow range of probabilities, i.e., just those below the threshold merge/deduplicate. For example, if the automated program declares as duplicates records that match above 90% probability and non duplicates records that match below 70%, perhaps the registry decides to review matches between 85% and 90% instead of the full range of records that match between 70% and 90%.

4. During manual review, reviewers may identify aspects of the immunization events they are examining that are not considered by the automated algorithm. If these variables are systematic enough and clear enough, the registry may consider incorporating them into the automated deduplication system, thus resulting in a reduction in manual review.

5. Remaining records that cannot be resolved during the manual review should be considered duplicates (see BR09, p. 22) and handled according to procedures established in the IIS (keep, discard, suppress, etc).

False matches (false positives) and false non-matches (false negatives)

Terms for discussion:
True match: Two vaccine entries that are duplicates are identified as duplicates by the IIS
False match (false-positive): Two vaccine entries are NOT duplicates, but are identified as duplicates by the IIS.

True non-match: Two vaccine entries are NOT duplicates and the IIS identifies them as non-duplicates.
False non-match (false-negative): Two vaccine entries are duplicates, but the IIS identifies them as non-duplicates.

The following describes potential scenarios and their results regarding the deduplication process:

1) Deleting/suppressing a true match from the IIS → fine. Two vaccine entries that are true duplicates are identified as such by the IIS and one is deleted/suppressed. This would be the optimum action to take if a duplicate record is identified.
2) Keeping a false non-match in the IIS → could result in underimmunization of the child because it results in having duplicate vaccines recorded. If the duplicate vaccine is kept in the IIS, and counted as a legitimate vaccine/dose, the child looks as if he/she received this dose, but really did not.
3) Deleting a false match in the IIS → could result in overimmunization of the child because one of the vaccine entries could be suppressed or deleted when it should not be. If this true non-duplicate shot is deleted from the IIS and/or no longer viewed as part of the child's record, a health care provider may then vaccinate the child again for the same vaccine.
4) Keeping a true non-match → fine. Two vaccine entries that are not duplicates are identified as not being duplicates and become two separate entries.

	Match	Non-Match
Duplicates	True match (1)	False non-match (2)
Non-duplicates	False match (3)	True non-match (4)

When making a determination regarding a possible vaccine duplicate, always err on the side of a child being overimmunized as opposed to underimmunized (i.e., the child should receive more than the number of recommended vaccines as opposed to less than the number of recommended vaccines – see BR09, p. 22). The Advisory Committee on Immunization Practices and the American Academy of Family Physicians published joint recommendations in 2002 that support vaccination when documentation is in doubt in order to reduce the number of susceptible persons in the population.

"*Vaccination providers frequently encounter persons who do not have adequate documentation of vaccinations. Providers should only accept written, dated records as evidence of vaccination. With the exception of pneumococcal polysaccharide vaccine, self-reported doses of vaccine without written documentation should not be accepted. Although vaccinations should not be postponed if records cannot be found, an attempt to locate missing records should be made by contacting previous health-care providers and searching for a personally held record. If records cannot be located, these persons should be considered susceptible and should be started on the age-appropriate vaccination schedule. Serologic testing for immunity is an alternative to vaccination for certain antigens (e.g., measles, mumps, rubella, varicella, tetanus, diphtheria, hepatitis A, hepatitis B, and poliovirus) (see Vaccination of Internationally Adopted Children).*"

http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5102a1.htm

Note regarding the impact of the date window's size
Because of the size of the date window (23 days maximum) recommended for use during the Selection phase of the process, false-negatives (false non-matches) would not lead to under-immunization and false-positives (false matches) would not lead to overimmunization since all vaccinations under consideration are no more than 23 days apart (28 days interval between shots —minus 4 days grace period—is necessary in order for the shot to be considered valid for the routine vaccines).

If two truly duplicate records are evaluated as non-duplicates (false non-match), then a record is added to the registry, but it would be considered invalid since it is within 23 days or less from another similar shot and child would still receive a vaccination if it is necessary according with the schedule (no underimmunization occurs as a result of deduplication).

If two different records are evaluated as duplicates (false match), then a record will not be added to the registry, but the outcome is the same since the original shot (if added) would be invalidated anyway because it is within 23 days or less from another similar shot and child would receive an additional/replacement shot if it is necessary according to the schedule (no overimmunization occurs as a result of deduplication). At the same time this type of error, i.e., false match, leads to a loss of data.

So, operational level decisions made during the deduplication process (false non-match and false-match) would not affect programmatic level outcomes (underimmunization, overimmunization) because of the selected size of the date window, 23 days maximum.

Table 15. Matrix of evaluation outcomes for two records (23 days apart or less)

	Condition	**Record A: evaluation outcome**	**Record B: evaluation outcome**	**Record B: Correct (should-be) evaluation outcome**
1	True match	Added	Suppressed / deleted	Suppressed / deleted
2	False non-match (false negative)	Added	Added; invalidated (<28 days) (***wrong data***)	Suppressed / deleted
3	False match (false positive)	Added	Suppressed / deleted (***loss of data***)	Added; invalidated (<28 days)
4	True non-match	Added	Added; invalidated (<28 days)	Added; invalidated (<28 days)

Chapter 7: Conclusions

The consistent use and implementation of these guidelines will help improve the vaccination level deduplication practices in immunization information systems and help ensure that accurate clinical decisions will be made. The following summary is a brief description of the key outcomes and accomplishments of this workgroup.

Major consensus-based accomplishments of the workgroup:

- Defined vaccination level deduplication process as a three-phase process that includes Selection, Evaluation, and Resolution phases.
- Developed principles that vaccination level deduplication process should be based on, business rules to follow, and specific scenarios that illustrate application of principles and business rules.
- Formulated a concept that each vaccination event should be presented in the registry by two records (views) – the best record (view) selected from vaccination reports (utilized for clinical purposes) and the consolidated record (view) that aggregates information from all available sources (utilized for public health and clinical purposes).
- Created specific detailed recommendations for selecting the best record and for constructing the consolidated record.
- Developed a decision matrix for recommendations on levels of confidence based on sources of information.
- Developed and reconfirmed key definitions for the topic of vaccination level deduplication.

Primary process acknowledgments from the workgroup's activities:

- In spite of differences in immunization registry programs, common approaches can be discovered and agreed upon.
- Business modeling initiative provides an efficient venue for collaboration and exchange of ideas among peers.
- Business modeling and facilitation techniques help to reach a consensus and document agreed upon approaches.
- Facilitated sessions promote and organize brainstorming and allow groups to achieve results.
- Business modeling promotes and organizes the analysis and improvement of immunization operations.

Presented approach and results are relevant for and can be used beyond immunization information systems—for developing and documenting best practices and operational requirements for domain-specific deduplication applications in public health, healthcare, and other areas.

The results of this project are intended to support a uniform alignment of the vaccination level deduplication processes in immunization registries according to recommended guidelines.

References

1. Miller PL, Frawley SJ, Sayward FG. IMM/Scrub: a domain-specific tool for the deduplication of vaccination history records in childhood immunization registries. Computers and Biomedical Research 2000;33:126–143. Abstract available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=10854120&dopt=Abstract
2. Salkowitz SM, Clyde S. De-duplication technology and practices for integrated child-health information systems. Decatur, GA: All Kids Count, Public Health Informatics Institute, 2003. Available at http://www.phii.org/Files/dedupe.pdf
3. Public Health Informatics Institute. The unique records portfolio. Decatur, GA: Public Health Informatics Institute, 2006.
4. Angeloni M, Berry M. Probabilistic record matching and deduplication using open source software. Presented at the Immunization Registry Conference, Atlanta, GA, October 2004. Slides available at http://www.hln.com/assets/pdf/IRC04-KIDSNET.pdf
5. Davidson K. The deduplication numbers game. White paper. Greenville, SC: QS Technologies. Available at http://www.qsinc.com/about/papers/dedup.html
6. Borthwick A, Soffer M. Business requirements of a record matching system. ChoiceMaker Technologies, Inc. Available at http://www.choicemaker.com/content/publications/index.php3
7. Borthwick A. Evolution of user requirements for deploying an advanced record matching solution. Presented at the Immunization Registry Conference, Atlanta, GA. October 28, 2003. Available at http://www.choicemaker.com/content/publications/index.php3
8. Borthwick A, Papadouka V, Walker D. Principles and results of the NY Citywide Immunization Registry's MEDD De-Duplication Project. Presented at the 34th National Immunization Conference. Washington, DC, July 7, 2000. Available at: http://www.choicemaker.com/content/publications/index.php3
9. Centers for Disease Control and Prevention. National Immunization Survey (NIS) Reference Manual. Atlanta: CDC.

Appendix A: Domain model

Domain model purpose and explanation

A domain is an area of knowledge or activity characterized by a set of concepts and terminology understood by the business practitioners in the area.

A domain model captures a business vocabulary—terms and definitions. It ensures that all terminology and concepts that will appear in the process description and business rules are known and understood by the domain practitioners (agreed-upon definitions and meaning).
A domain model includes:

- A domain diagram that shows major business entities, their relationships and responsibilities (**Fig. A-1**).
- A table of entities and attributes that provides the full descriptive details of the components represented on the diagram (**Table A-1**)

Entities and attributes on Fig. A-1 are numbered; these numbers correspond to numbering of rows in Table A-1 where entities and attributes are described.

Unlike a data model diagram which depicts storage of information, or a workflow/process diagram which depicts the sequence of steps in a process, a domain diagram is a high-level static representation of the main "things" (entities) involved in the immunization process, including a description of how these "things" (entities) are related. It is important to note that the domain diagram is not a technical specification. Instead, the domain diagram provides the foundation for other modeling diagrams and materials.

How to read and interpret the domain diagram:

- Relationships between entities are visualized by connecting lines.

- Names associated with these lines describe the type of the relationship between entities. Example: a relationship between *Vaccine* and *Antigen* is shown as a connecting line with the word (label) "contains". Such a relationship should be read as "*Vaccine* contains *Antigen*".

- The general convention for interpretation of relationships between entities is to construct such a description by reading clockwise, starting from the first entity name (*Vaccine*), then relationship name—contains (note, that the name is shown on the top of the line, supporting a clockwise reading), then the second entity name (*Antigen*).

- If we need to read the same description in the opposite direction, from *Antigen* to *Vaccine*, we would have to place a second name— "is contained in" —below the line. In this case, using the clockwise reading rule, a description would be "*Antigen* is contained in *Vaccine.*" In most cases just one name for a relationship is employed (like "contains" in the example just considered) assuming that it should be sufficient for a proper interpretation of a relationship in both directions.

Description of the domain diagram

The entities and their characteristics (attributes) presented on the domain diagram **(Fig. A-1)** describe a limited fragment of the immunization domain related to the vaccination level deduplication. Key entities for the vaccination level deduplication are Vaccination Event, Vaccine, Vaccination Event Submission (Report), and Submitter (the source of information about a vaccination event); details on all of these and other entities are presented in **Table A-1**.

Patient is getting vaccinated as a result of the ***Vaccination Event***. More than one Vaccination Event can happen during the Vaccination Encounter (office visit). In other words, Patient can receive several vaccine shots during a single office visit; each shot would be represented by a dedicated vaccination event. Accordingly, the relationship between Vaccination Event and Vaccination Encounter is labeled with "1" for the Vaccination Encounter and "1…n" (meaning one or many) for the Vaccination Event.

Vaccine refers to a product that produces an immune response in a patient and is administered during the Vaccination Event. It is described by a set of characteristics (attributes), such as vaccine type, CVX code, CPT code, trade name, lot number, etc. A single Vaccine can be related to multiple Antigens, such as tetanus, diphtheria, pertussis.

Information about the Vaccination Event comes to a registry from a ***Submitter*** **(**also known as a Source) in the form of a ***Vaccination Event Submission (Report).*** Typical Submitters would be a vaccination Provider Organization and medical insurance company. Characteristics (attributes) of the Submitter and the Vaccination Event Submission (Report) are important in evaluation of vaccination data for deduplication purposes: there are different levels of confidence in different sources and ways they report regarding the quality of their data. For example, documentation Type and Method of submission affects the level of confidence in the immunization data (e.g., direct entry to the registry vs. electronic feed).

Vaccination Event Record (not shown on the domain diagram) includes information about the vaccination event from all domain entities.

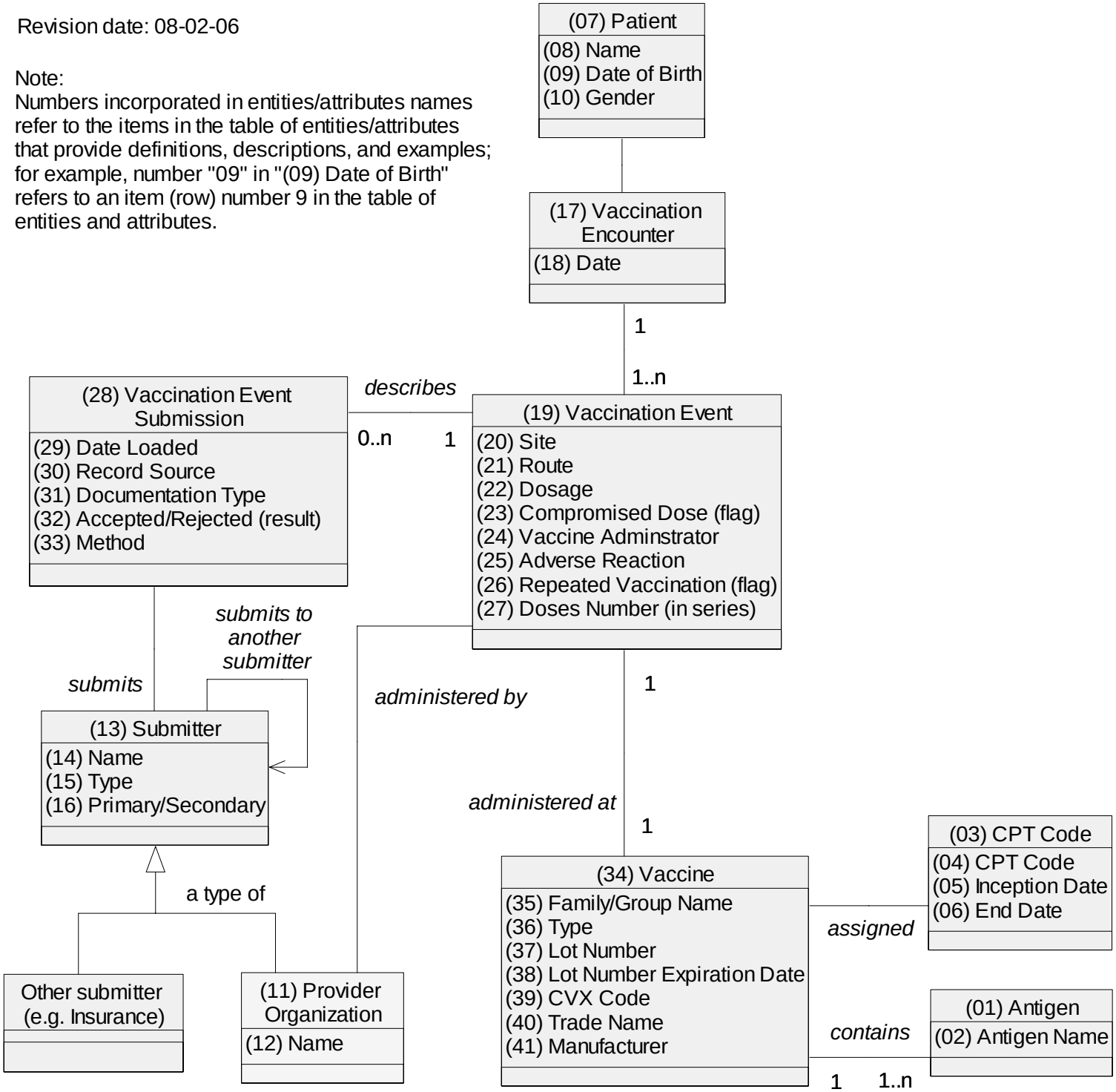


Figure A-1. Domain diagram for vaccination level

Table A-1. Entities and attributes (terms and definitions)

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
1.	**Antigen** [2]		Antigen: a foreign (non-self) substance (microbe) found in the body that can be either live (such as viruses and bacteria) or inactivated proteins and polysaccarides. Vaccinations allow the immune system to develop a defense against antigens. Every vaccine relates to one or more antigens.	See CDC "Epidemiology and Prevention of Vaccine-Preventable Diseases, " 9th edition, Jan. 2006, Chapter 1- Principles of Vaccination. Also known as the "Pink Book" Vaccines are designed to confer immunity against specific disease antigens or toxins, like measles, polio and diphtheria. One or more doses of a vaccine, administered over a period of time, may be required to produce long-lasting immunity. Information about antigens usually derived from other immunization information, e.g. vaccine type, trade name.
2.	Antigen	Name		Examples: diphtheria, tetanus, pertussis. Usually derived from other immunization information, e.g. vaccine type, trade name HL7: n/a CDC Core Data Set: N
3.	**CPT Code**		Procedure Code related to the vaccine.	These are billing codes. Used along with CVX codes (see below). Some CPT Codes have been re-used. There are vaccines that do not have CPT codes. CPT™ codes from the "Immune Globulins and Vaccines, Toxoids" sections of the American Medical Association's Current Procedural Terminology (CPT) 2006 http://www.cdc.gov/nip/registry/st_terr/tech/stds/cpt.htm

[1] Formatting : a) for HL7: <name>(table/item number); b) for CDC Core Data Set: Y <name> – in data set, N – not in data set, N/A – not applicable.
[2] Rows related to Entities are shaded, rows related to Attributes of these Entities are not shaded.

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
4.	CPT Code	CPT Code	CPT: Stands for Current Procedural Terminology. Codes developed by the American Medical Association that stand for medical or psychiatric procedures performed by health care practitioners	Example: 90700 – DTaP AMA: http://www.ama-assn.org/ama/pub/category/3113.html HL7: Administered Code (0292/00347) CDC Core Data Set: N
5.	CPT Code	Inception Date	The inception date is the published date when the CPT code is allowed to be used The inception date may not be the date of licensure of the vaccine	HL7: n/a CDC Core Data Set: N
6.	CPT Code	End Date	The end date is when the published CPT code is not allowed to be used for that vaccine	HL7: n/a CDC Core Data Set: N CPT codes are typically valid for certain periods, therefore dates can be used to assess validity.
7.	**Patient**			While patient/client deduplication is crucial, it precedes immunization deduplication and may not be needed in this domain model beyond indicating that we have a unique person. Accordingly, just a few attributes for the Patient are present in this domain model.
8.	Patient	Name		HL7: Name Type (0200 /00108) CDC Core Data Set: Y (Patient name: first, middle, last)
9.	Patient	Date of Birth	The date of birth either stated or reported on the patients' birth.	Used for data validation, e.g. vaccinations given before birth date are invalid. HL7: Date of Birth (00110) CDC Core Data Set: Y (Patient birth date)
10.	Patient	Gender	This is the observed or reported patient's Sex.	One vaccine at this time is gender specific—HPV for females between the ages of 9 and 26. HL7: Sex (0001/00111) CDC Core Data Set: Y (Patient sex)

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
11.	**Provider Organization**		A provider organization is a collection of related providers (clinicians) that are treated as an entity that administered (owns) immunizations.	The provider organization of the vaccination event may be unknown
12.	Provider Organization	Name	This name can be a corporate name and may include a number of different provider offices/sites and physician groups.	Example: Dr. Smith Associates HL7: Administered-at location (0292/00353) CDC Core Data Set: Y (Vaccine provider)
13.	**Submitter**		Entity that submits immunization data to the registry	Also known as source – indicates where the information came from.
14.	Submitter	Name	Name of the organization that submits a report regarding the vaccination event.	HL7: n/a CDC Core Data Set: N

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
15.	Submitter	Type	Can include all types of providers and organizations submitting data to the registry: Values can include but are not limited to: o Public and private healthcare providers (administering vaccines) o Public and private healthcare providers-“ Other” (entering historical shot info) o Health plans – 3rd Party Provider or HMO o Public and private schools - administering shots and/or entering historical shot information o Other registries—by electronic interfaces or batches o WIC-entering historical information o Head Start centers and child care providers o Medicaid -billing batches	Insurance is one type of submitter See also Primary/Secondary for Vaccination Event Submission HL7: see table NIP007 CDC Core Data Set: N

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
16.	Submitter	Primary/Secondary	The "primary submitter" is the PROVIDER ORGANIZATION that submits a VACCINATION EVENT SUBMISSION and claims that it administered the VACCINATION EVENT. The "secondary submitter" is a SUBMITTER that submits a VACCINATION EVENT SUBMISSION and is not the PROVIDER ORGANIZATION that administered the VACCINATION EVENT.	A secondary submitter data can be differentiated from the primary submitter if the administering provider name is included in the data for each vaccine dose. HL7: Sending Facility 00004) or NIP 001 Immunization Information source? CDC Core Data Set: N
17.	**Vaccination Encounter**		The visit data	Several vaccination events can happen within one vaccination encounter.
18.	Vaccination Encounter	Date	Date of the vaccination. The date that the patient received dose(s) of the vaccine(s)	HL7: Date/time start of administration (0292/00345) or Date/time end of administration (0292/00346) CDC Core Data Set: Y (Vaccination date)
19.	**Vaccination Event**		Administration of one Vaccine to a Patient. Several vaccination events can happen during one vaccination encounter.	Example: a "needle stick", a tablet The presence or absence of some of these items may inform the decision to determine the best immunization from a match.
20.	Vaccination Event	Site	Anatomical site where immunization was administered.	CDC Pink Book: Appendix D: Vaccine Administration Guidelines, page 5:Administering Vaccines- Dose, Route, Site, and Needle Size HL7: Administrative site (0163/00310) CDC Core Data Set: Y (Vaccine injection site)
21.	Vaccination Event	Route	The method of administration. Includes: injection, intranasal or oral.	HL7: Administrative Route (0162/00309) CDC Core Data Set: N

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
22.	Vaccination Event	Dosage	The measurement of how much vaccine was administered. The measurement can be as a pediatric or adult dose and counted as 1, 2, 3, etc. or in actual centiliters (cc's).	HL7: Administered Amount (0292/00348), Administered Units (0292/00349) CDC Core Data Set: N
23.	Vaccination Event	Compromised Dose (flag)	A flag indicating that a dose of vaccine should not considered when evaluating the immunization history. Valid/Invalid indicator. Indicates that a dose administered to a patient is considered substandard and therefore not a valid dose.	There are a number of reasons that a dose of vaccine is not considered when evaluating the immunization history. These include: -immunization with product that has been compromised (too old, over-heated, or expired), -uncertain quantity of vaccine has been administered (e.g., the child jerked away during administration and some of the vaccine was spilled). Recording them in the registry is important for several reasons including: - maintaining accurate inventory; - noting lots used (in case of recall). - in some cases a repeat dose is given immediately. Deduplication algorithms should not identify these as duplicate records, but rather duplicate vaccines of which the 2nd dose is the valid (kept) dose. HL7: n/a CDC Core Data Set: N
24.	Vaccination Event	Vaccine Administrator	The person who gave the vaccination (as opposed to the Provider Organization above)	Example: Dr. Smith HL7: Administering Provider (0292/00352) CDC Core Data Set: n/a

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
25.	Vaccination Event	Adverse Reaction	Used to indicate if an adverse event (AE) is associated with a vaccination event A vaccine adverse event can include any adverse reaction that the patient experiences after receiving a dose of vaccine.	See http://www.vaers.hhs.gov/reportable.htm HL7: n/a CDC Core Data Set: N
26.	Vaccination Event	Repeated Vaccination (flag)	A flag indicating that a dose of vaccine has been administered as a remedy to a compromised dose (see above)	HL7: n/a CDC Core Data Set: N
27.	Vaccination Event	Doses Number (in series)		HL7: n/a CDC Core Data Set: Y (Vaccine dose number)
28.	**Vaccination Event Submission**		A report regarding the vaccination event	A submission could be sent twice (or more) - once by provider organization and once by other source
29.	Vaccination Event Submission	Date Loaded	The date that the vaccine event data were loaded into the registry (usually from a batch file). This is the completion of the data submission.	This data may not be visible to the user from the GUI (graphic user interface) and therefore may not be available. Also, this date should not be confused with the vaccine event date. This is not necessarily the date that the report was sent. This is not necessarily the date the report was received. HL7: n/a; Internal variable used to monitor when data is ready to take action. CDC Core Data Set: N

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
30.	Vaccination Event Submission	Record Source Type	Values: Administered or Historical. “Administered” means that submitter attests that (s)he gave this shot (administered the vaccination event). All other cases are considered to be “Historical”.	“Historical” is another provider’s data. "Historical" – can come only from Secondary submitter; also, "Transcribed" (see below) can come only from Secondary submitter. "Administered" can come only from Primary submitter. Possible situation: more than one submitter claims the shot. HL7: NIP 001 Immunization Information source? or Source of comment (0105/ 00097) ? CDC Core Data Set: N
31.	Vaccination Event Submission	Documentation Type	Values: o Billing/Claim o Clinical o Transcribed Transcribed describes information received from the "secondary submitter", e.g., second hand information.	Example: one doctor called to another doctor for the information. HL7: n/a CDC Core Data Set: N
32.	Vaccination Event Submission	Accepted/ Rejected (the result)	An accepted vaccine event submission means that the data conformed to the data guidelines and was or will be uploaded into the registry. A rejected vaccine event submission means that the vaccine event data did not conform to the data submission guidelines of the registry and was or would not be uploaded into the registry.	HL7: n/a CDC Core Data Set: N

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
33.	Vaccination Event Submission	Method	This has nothing to do with the information content, it is strictly how the information was submitted (the medium). Values o Electronic Interface (HL7 or other) o Registry Specific User Interface	User interface is an opportunity to seek human evaluation and decision. Some registries may use this information to direct them how to evaluate duplicates. They may trust data via the user interface more than electronic or batch data. User interface, although not entirely error-free, is an opportunity for human evaluation and decision. HL7: n/a CDC Core Data Set: N
34.	**Vaccine**		A vaccine is a product that produces an immune response in a patient. The product that is given; the VACCINE that was administered at the vaccination event	Vaccines can be identified generically or by brand name (or both) in a registry. Some registries identify equivalent vaccines Issue: Antiviral (?) Example: The specific package in the refrigerator Absence / presence of vaccine data items is a criterion when evaluating which record is best.

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
35.	Vaccine	Family/ Group Name	The ACIP Immunization Schedule describes several categories, which in general correspond to individual antigens, and sets forth, for each one, the number of immunization doses that will be required to produce complete immunity, based on immunological research and clinical trials. These categories will be referred to as immunization series (*Families/Groups*). http://www.cdc.gov/nip/registry/st_terr/tech/peg/prog_eval_guide.pdf The Vaccine Family/Group Name is a broad category of vaccines that are related by vaccine type. There is not necessarily a one-to-one relationship between Family/Group Name and Type. There may be multiple family/group names for one vaccine.	ACIP classification of the Vaccine Family/Group Name should be followed as the standard. It may be extended as necessary Example: HIB, HepB, IPV Examples are: The DTP family can include: DT, DTaP, Tdap, Td, Td. Although immunization series (Family/Group) in the ACIP Schedule generally corresponds to single antigens, this is not always so: DTP (diphtheria-tetanus-pertussis) and MMR (measles-mumps-rubella) are regarded as series (Family/Group), even though they each refer to multiple antigens. It would be more to the point to say that immunization series (Family/Group) correspond to vaccines, rather than antigens, since both DTP and MMR are vaccine products. http://www.cdc.gov/nip/registry/st_terr/tech/peg/prog_eval_guide.pdf Crucial in the first step of finding pairs to consider. Family/Group name usually derived from other information, e.g. trade name, CVX code, etc. HL7: n/a CDC Core Data Set: n/a

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
36.	Vaccine	Type	A numerical code designating vaccine administered. Codes can indicate antigens, route of administration, dose schedule and age formulation. The vaccine type is the specific type of vaccine(s) given during the vaccine event.	Examples: HIB-HBOC, HIB-HepB This should map to CVX. There is normally one CVX Code per one vaccine type. There are vaccine types that do not have a CVX code. The vaccine type can include single types of vaccines as well as combination vaccines, e.g., IPV, or IPV, DTaP, HepB B. Also the vaccine type data may indicate a generic or specific type of vaccine: e.g. pneumo or PCV7 or PPV23 HL7: Administered Code (0292/00347) – this is CVX code CDC Core Data Set: Y (Vaccine type)
37.	Vaccine	Lot Number	The lot number is the number assigned by the manufacturer for a specific batch of vaccine developed and distributed. This is the tracking number of the administered vaccine.	HL7: Substance Lot Number (0292/01129) CDC Core Data Set: Y (Vaccine lot number)
38.	Vaccine	Lot Number Expiration Date	Manufacturers are required to assign a lot number expiration date to each batch of vaccine.	Note: This is important to some registries to provide reasons why a record was not loaded (not all registries have business rules associated with this) HL7: Substance Expiration Date (0292/01130) CDC Core Data Set: Vaccine expiration date

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
39.	Vaccine	CVX Code	The typical representation of the vaccine type reflected as a CVX code. There is generally one CVX code per one vaccine type which may include one or more vaccines.	Example: 47 – HIB-HBOC CDC assigns specific CVX codes to each vaccine or combination of vaccines. Note: there are CVX codes for vaccines that do not exist. Those codes are considered placeholders. There have been issues in the past where some have been reassigned but generally this is not supposed to happen There are vaccine types that do not have a CVX code. It was intended that a CVX code would be paired with the code for Manufacturer (s) (code = MVX) and this pair would indicate a specific vaccine, e.g. brand name. HL7: Administered Code (0292/00347) CDC Core Data Set: Y (Vaccine type)
40.	Vaccine	Trade Name	The Trade Name is the name under which the manufacturer copyrights the vaccine(s). Trade name is synonymous with the Brand name. A trade name usually assigned by manufacturer to identify vaccine type.	Example: ACTHIB, Comvax CDC Pink Book, Appendix B, U.S. Vaccines http://www.cdc.gov/nip/publications/pink/appendices/B/us-vaccines.pdf If Trade Name is not actively collected by a particular IIS, it can be derived from other variables, e.g., Vaccine Type and Manufacturer Name. HL7: n/a CDC Core Data Set: N

	Name		Description	Remarks & Relationship to other standards (HL7) [1]
41.	Vaccine	Manufacturer	Organization that manufactured a vaccine. The manufacturer refers to the company that develops and distributes vaccine(s). There may be several manufacturers of a particular vaccine family/group and/or vaccine type.	CDC assigns an MVX code to specific vaccine manufacturers. Issue: The manufacturer can change over time when a brand is sold to a different manufacturer. It was intended that a MVX code would be paired with the code for vaccine (code = CVX) and this pair would indicate a specific vaccine, e.g. brand name. HL7: Substance manufacturer name (0227/01131) CDC Core Data Set: Y (Vaccine Manufacturer) HL7: n/a CDC Core Data Set: N

Appendix B: Merging Data From Duplicate Records

Issue in brief
On one hand, clinical records in many states can not be changed by anybody but the owner; on the other hand, records have to be changed to aggregate multiple pieces of information (adverse events, administrative data, lot numbers from inventory sources etc) from duplicate records and other sources. That controversial requirements lead to a strategy of having two primary sources of information for each vaccination event as a way to satisfy both sides of these requirements.

Background
Currently most of the registries (except Michigan) do not routinely merge information from the duplicate records; rather they select the best record among duplicates to represent a vaccination event. Because of restrictions on changing clinical records (only owners can do that) the additional information about a vaccination event, available from other duplicate records and other sources, is not used to enhance the selected best record, to make it more complete.

While such an approach serves relatively well to address individual clinical functions of an IIS, it does not provide the best basis for maintaining the most complete record of the vaccination event and performing various functions of the IIS. From a public health perspective, a registry has to aggregate all available information about a vaccination event into the consolidated record. For example, it should combine the information from provider who administered the shot (in most cases this is the best clinical record) with information from other duplicate records and other sources—such information as lot number, vaccine trade name, or vaccination adverse reactions.

A possible approach would be for a registry to maintain two records (or views) for each vaccination event - first, the selected best record for clinical purposes, and second, the merged best record for public health and clinical purposes. This consolidated or composite immunization record could provide specific clinical information needs, such as being able to associate a lot number in a recall effort with an individual.

Example
Let's say that we have two duplicate records - record A and record B (see **Figure B-1**) - that describe the same vaccination event. Assume that both records are from trusted sources and contain good data. Record A contains data items x (vx type) and y (vx trade name); record B contains data items x (vx type) and z (vx lot #). We can present these duplicate records in a form A[x,y] and B[x,z]; obviously, this is a simplified example and real records have more than just two variables. What record should represent a vaccination event in this case?

The first possible approach would be to select the best record to represent a vaccination event; the second possible approach would be to create a new consolidated record C that aggregates all available information from record A[x,y] and record B[x,z]. The Pro and Con descriptions for each approach are presented below. A reasonable combination of both approaches would be to have one best record - record A[x,y] for clinical purposes, and second record - record C[x,y,z] - for public health and clinical purposes.

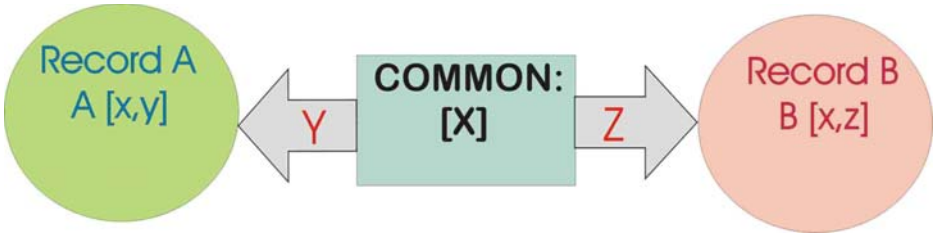


Figure B-1. Illustration / example for the consolidation/composite record issue

Possible choices are to either merge information or select the best record

Example - Best record approach
The approach of selecting the best record to represent a vaccination event is illustrated with **Figure B-2** and **Table B-1.** Let's say that record A[x,y] is selected as a best record. No information from the record B[x,z] is used; specifically we are not using available data item z (vx lot #) from the record B in representation of the vaccination event. But such an approach satisfies the restriction that clinical records can not be changed by the IIS.

Table B-1. Pros and cons for the best record approach

Pros	*Cons*
Saved '*original*' information from provider and that's why contains high validity data.	Lead to loss of additional information about important Immunization Vaccine Identifiers (e.g. lot #).
Satisfies conditions of clinical records requirements.	Lose opportunity to create maximal knowledge about the vaccine (encounter) from all sources, such as vaccine adverse events.
Provides cost benefits (no additional resources for manual review and merging).	Limit growth by restricting the linkage of other information to the event.
Good for clinical, legal and audit reasons.	

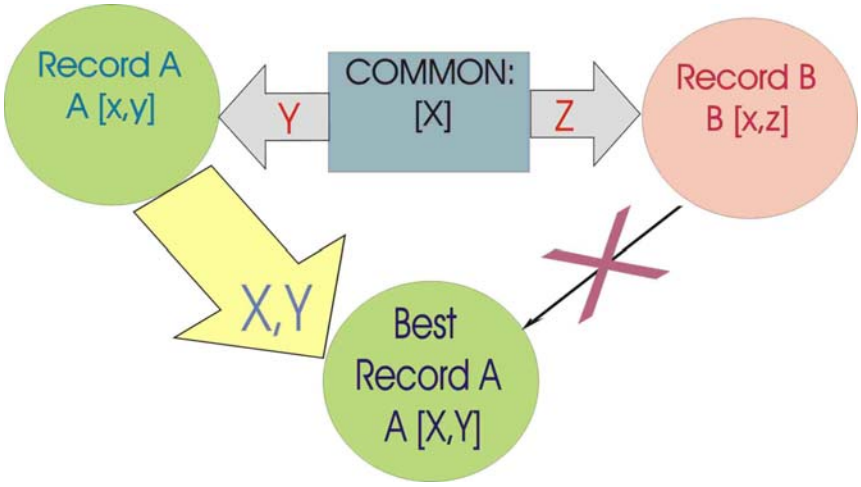


Figure B-2. Best record approach - for clinical purposes

Example - Consolidated record approach
The approach of creating a new consolidated record C that aggregates all available information from record A[x,y] and record B[x,z] is illustrated with **Figure B-3** and **Table B-2**. Resulting record C will contain all available to us data items x, y, and z - C[x,y,z]. Such an approach satisfies requirements for public health records to present aggregated information from all available sources.

Table B-2. Pros and cons for consolidated record approach

Pros	*Cons*
Contains combined information from different sources and provides more complete data about vaccination event.	More resources consumed to ensure quality is maintained or improved.
Supports better analysis with more complete data.	Business rules become more complex for consolidated record.
Increases likelihood for a unique match (more additional information provided).	Increases risk of errors occurring in data processing.

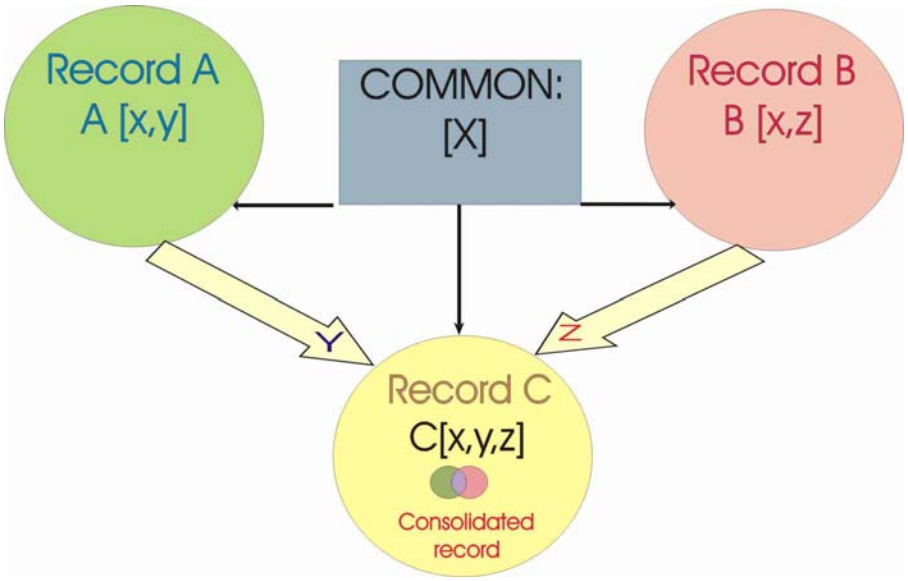


Figure B-3. Consolidated record approach - for public health purposes and clinical purposes.

Can consolidation of information for a vaccination event be justified?
Some subject matter experts feel that the creation of a consolidated record is unwarranted for two reasons: a) simplicity would be lost; b) not sure that it would really benefit public health functions in a significant way (in other words, that the additional information that consolidated record has compare to the best record would make that much of a difference). Following data sample from the Minnesota IIS provides some quantitative view on this issue:

DTaP Analysis
85 possibly duplicate randomly selected DTaP vaccinations (denominator);
38 needed manual review (various reasons, no useful information to add);
39 had nothing of value in terms of additional information;
8 had good additional information (2 had historical versus administered indicators, 2 had lot numbers the other didn't have, and 4 had trade names).
8/85 had useful information (so that is 9.4%).

HepB Analysis
An examination of possible duplicate HepB's given from 1/1/06-8/1/06 was conducted. A total of 62,785 shots given, there were 138 sets to analyze (sets usually meaning 2 shots but in some cases there were 3). This analysis shows so far 20.2% from this HepB analysis would have benefited from a merge.

Provider same/diff:
40 sets had different providers
98 sets were from the same providers

Admin/Hist:
23 sets were administered
90 were historical
25 were a mix of both

Batch loads/Direct Data entry:
17 sets were from direct data entry
97 were from batch loads
13 were a mix of both

Vaccine Type:
112 sets had same vaccine type
26 had different vaccine type

Trade Name:
90 sets had no trade names
27 had trade names that matched
1 had a trade name that was different
20 had trade names that would be additional info

Lot numbers:
86 sets had no lot number
2 were obvious typos
15 had lot numbers the same
7 had different lot numbers (These 7 were also from the same provider.)

Recommendations
Because there are a variety of processes involved in IIS (e.g. clinical, public health, legal), it is important to recognize that additions and deletions to vaccination records affect stakeholders differently. For example, a clinician may not want one of their vaccination records deleted or merged with an entry provided by another source, even if it is a duplicate, because they have a vested interest in vaccinations administered in their offices. A public health agency, though, may prefer that duplicate vaccinations be deleted or merged with other records as appropriate, because their main interests are the vaccination levels of the entire community. Furthermore, there may be legal issues involved in states regarding deleting or merging records. Thus, to ensure that the needs of these multiple stakeholders are met, workgroup recommends the following regarding duplicate vaccinations in the IIS:

> Allow all vaccination entries into the IIS database. Within the entire IIS create 2 screens for viewing the following:
>
> View 1— vaccines administered view. Includes all vaccines given at a particular provider's office and allows viewing and editing of these vaccines by appropriate office personnel. An electronic version of the immunizations should be displayed in a patient's medical chart.
>
> View 2 – composite view. Includes all vaccine entries from all sources and allows their viewing by appropriate health officials. Vaccines are deduplicated and under quality assurance. Data that populate this view are used in forecasting algorithms, analysis reports and are printed on official records.

The creation and adoption of the consolidated record may require some significant retooling of current operations and technical capacity of the registry, but ultimately the creation of a consolidation record at the vaccination level is the only way to ensure that the immunization event is documented most comprehensively into a single, accurate immunization record. Therefore adopting the strategy to create a consolidated record will more precisely address the intent of the NVAC functional standard #12, which as stated is to promote accuracy and completeness of registry data by using a data quality protocol to combine all available information into a single, accurate immunization record.