

**DEVELOPMENT OF AN EPIDEMIOLOGIC DATA WAREHOUSE  
FOR THE OREGON IMMUNIZATION ALERT REGISTRY**

**By**

**Neal D. Goldstein**

**A CAPSTONE PROJECT**

Presented to the Department of Medical Informatics & Clinical Epidemiology  
And the Oregon Health & Science University  
School of Medicine

In partial fulfillment of  
The requirements for the degree of

**Master of Biomedical Informatics**

**December 2006**

School of Medicine  
Oregon Health & Science University

Master of Biomedical Informatics

---

**CERTIFICATE OF APPROVAL**

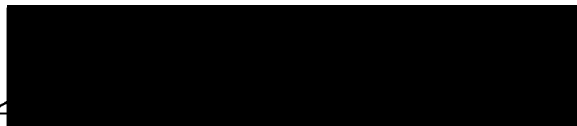
---

This is to certify that the Capstone Project of

**Neal D. Goldstein**

*"Development of an Epidemiologic Data Warehouse for the Oregon Immunization  
ALERT Registry"*

Has been approved



Judith R. Logan, MD, MS

12-21-06  
Date

## Table of Contents

Table of Contents .....	i
List of Tables .....	iii
List of Figures .....	iv
Acknowledgements .....	v
Abstract .....	vi
Introduction .....	1
Overview of Electronic Immunization Registries .....	2
Benefits .....	3
Challenges .....	5
Data Quality .....	5
Participation .....	6
Economics .....	6
Integration .....	7
Performance .....	8
Current State of Immunization Registries .....	8
About the Oregon Immunization ALERT Registry .....	10
Architecture .....	10
Tables .....	11
Demographics .....	12
Phone .....	12
Address .....	12
AliasNames .....	13
Vaccinations .....	13
IDAlias .....	14
Table Relationships .....	14
Workflow .....	15
Secondary Data & Subsystems .....	16
Motivation for a Research Data Warehouse .....	18
Methodology .....	19
Sample Research Questions .....	19

Questionnaire .....	21
Data Model.....	25
Databases .....	25
Tables.....	26
Demographics .....	27
Vaccinations.....	27
TemporalData .....	27
IDsLookUp .....	28
Table Relationships.....	28
Extraction, Transformation, and Loading.....	34
Maintenance.....	36
Summary and Conclusion .....	37
References.....	39
Appendix A. ALERT Core Tables with Fields.....	42
Appendix B. ALERT Research Data Warehouse Questionnaire.....	43
Appendix C. Summary of Questionnaire Results.....	46

## **List of Tables**

Table 1. Examples of potential benefits of immunization registries. ....	4
Table 2. Challenges to the success of immunization registries. ....	5
Table 3. ALERT core tables. ....	11
Table 4. Secondary data sources and subsystems of ALERT.....	17
Table 5. Databases of birth cohorts.....	26
Table 6. Birth cohort database core tables. ....	26
Table 7. Databases of birth cohorts with archived year.....	36

## List of Figures

Figure 1. ALERT core tables.....	15
Figure 2. ALERT registry workflow. ....	16
Figure 3. Relationships between the birth cohort database tables. ....	29
Figure 4. Demographics linked to IDsLookUp.....	30
Figure 5. Demographics linked to Vaccinations.....	31
Figure 6. Demographics linked to TemporalData.....	32
Figure 7. Vaccination de-duplication.....	35

## **Acknowledgements**

Thank you to the many people who made this project possible. To my advisor Dr. Judith Logan for her much valued guidance and expertise. To Barbara Canavan, Martha Skiles, Donald Dumont, Steve Robison, and the ALERT staff for sponsoring and funding this project. To Andrea Ilg for her support of the OHSU DMICE graduate program and the distance learners. To Dr. William Hersh for his leadership and support of the program. And, many thanks to the entire DMICE faculty and staff.

Special thanks to my parents Robin and Jeffrey, brother Kevin, and sister-in-law Blake, for their support and encouragement throughout my studies.

## **Abstract**

The Oregon Immunization ALERT is an electronic immunization registry that tracks vaccinations primarily for the state's pediatric population. The ALERT system is optimized as a clinical tool for improving vaccination rates and coverage, and therefore is not in a preferred format for research studies. This paper examines the current state of immunization registries, the architecture of ALERT, and proposes a model for development of an epidemiologic data warehouse.

In order to propose a data model for the warehouse, the researchers needs were ascertained. A questionnaire submitted to ALERT researchers yielded the pertinent information required in the warehouse. A subsequent meeting with the ALERT staff provided a forum for reaching a consensus on the core features of the research system.

The epidemiologic data warehouse is a subset of ALERT and contains the minimally required data fields. Additionally, the system incorporates various internal and external secondary data sets, which are specific to researcher interests. The success of this research system hinges on accurately capturing the researcher needs and providing timely vaccination information.



## **Introduction**

Immunization registries are information systems that maintain vaccination records for a pediatric population within a geographic area. The goal for the creation of a registry is to improve immunization rates in children, and subsequently improve public health. Data is not only submitted to these systems, but also routinely retrieved and analyzed to evaluate immunization coverage and program efficacy. It is this data retrieval process that serves as the motivation for this project, the proposal of an epidemiologic data warehouse for the Oregon Immunization ALERT registry.

The Oregon Immunization ALERT is an electronic immunization registry containing vaccination data for over 1.2 million children in the state. There are approximately 27 million vaccinations recorded in ALERT, with 500 to 600 public and private providers submitting data. The data stored in the registry is used for epidemiologic research and to serve as the master community record of vaccinations. The goal of ALERT is to help raise immunization rates for children in Oregon.

The data stored in ALERT is not in optimal format for research projects. Researchers would like a separate, static source (i.e., snapshot) of immunization records for conducting their studies. This paper proposes a model for a data warehouse that will satisfy the researchers' requirements. Additionally presented here is an overview of immunization registries, an overview of ALERT, methodology used to identify researcher requirements, and recommendations on ETL (extraction, transformation, and loading) and on-going maintenance of the warehouse.

## **Overview of Electronic Immunization Registries**

Immunization registries are electronic data repositories that maintain immunization records for a pediatric population in a geographic region. The availability of electronic vaccination coverage allows registries to function primarily as clinical decision support (CDS) systems (1). In this context, a CDS can generate reminders to parents or practitioners that a child is due for vaccination, or has missed a scheduled dose. An added benefit of having a single, common source of vaccination data for each region is simplifying epidemiologic studies, such as monitoring high-risk areas. The spring 2006 mumps outbreak in the Midwest is one example of how a registry could be used to track vaccination coverage for school-age children and identify at-risk areas. Several challenges need to be addressed for widespread acceptance and success of immunization registries. These include data security, provider participation, integration with other information systems, and sustainable funding (2).

Immunization registries have enjoyed a history of public and private support. The U.S. Centers for Disease Control and Prevention (CDC) has been an active participant in the funding, design, and development of these systems. A key federal government health objective for 2010 is to increase registry participation to 95% of the eligible population of children less than six years old (3). The Initiative on Immunization Registries, led by the National Vaccine Advisory Committee, has guided former President Clinton's original charter to create an "integrated immunization registry system" (2). Perhaps the greatest push for immunization registries has come from privately funded grants provided by the Robert Wood Johnson Foundation (RWJF). Through the RWJF All Kids Count programs

(All Kids Count I: 1992 – 1997, All Kids Count II: 1998 – 2000), the Foundation provided funding for 40 state and local registries (4).

Motivation for the use of an immunization registry stems from the need to track vaccination coverage. Vaccinations are one of the best preventive tools available, and ensuring proper immunization is vital (5). Freeman and DeFriesse identify three factors supporting the need for electronic registries (6). The first is the ever-increasing number of antigens and changes to the vaccination schedule. As of 2006, there are 10 recommended childhood vaccines, with up to 25 scheduled doses (7). The second factor identified is the common but incorrect assumption that a child's vaccinations are up to date. Registries can provide real-time access to a child's vaccination status for both parents and providers. Scattered immunization records, due to children seeing multiple providers, is the final motivating factor. This record scattering presents a significant barrier, but also an opportunity, for accurate and up-to-date record keeping, particularly in rural areas (8,9).

## **Benefits**

Benefits from implementing an immunization registry are available to parents, providers, organizations, and the general population. For convenience, these benefits can be divided into two categories: improvement of health and administrative support. Table 1 lists some of the potential benefits.

**Table 1. Examples of potential benefits of immunization registries.<sup>1</sup>**

Category	Benefit
<b>Improvement of Health</b>	
<i>Decision Support</i>	Provide reminders and recalls for doses due or missed
<i>Well Being</i>	Ensure up-to-date immunization, avoid over immunization
<i>Epidemiology</i>	Track vaccination coverage for an area, monitor high-risk areas, prevent disease outbreaks
<b>Administrative Support</b>	
<i>Documentation</i>	Official vaccination documentation for school, camp, or day care
<i>Clerical</i>	Simplify paperwork, manage inventory
<i>Metrics</i>	Create managed-care reports (e.g., HEDIS)

Consolidating vaccination records for a population into a single source allows providers to improve vaccination coverage in their community, avoid duplicated and missed vaccinations, and improve practice efficiency (10). Reminders for vaccination and recalls for missed doses can be automatically generated and sent to both parents and providers. Coverage reports can be generated to indicate under- and over-vaccinated areas, aiming to control vaccine-preventable diseases. Lower operating costs can result from making practices more efficient by eliminating the need to manually retrieve immunization records for each child and reducing duplicate immunizations (11). Despite the documented benefits of using reminders and recalls to improve vaccination rates, a study performed in 2003 found this core feature of registries is underused in public and private practice (12).

<sup>1</sup> Adapted from the National Vaccine Advisory Committee report on Development of Community- and State-Based Immunization Registries: <http://www.cdc.gov/nip/registry/pubs/nvac.htm>

## Challenges

There are challenges whenever an information system containing medical data is instituted. While the potential benefits of such a system are known from the onset, the barriers to the successful use of a system are often discovered post-implementation. Table 2 categorizes these challenges and provides a brief description of each.

**Table 2. Challenges to the success of immunization registries.**

Category	Challenge
Data Quality	Ensure accuracy, completeness, security, and uniqueness (avoid duplicates) of standardized data
Participation	Motivate public and private providers to use the registry
Economics	Ensure public and private funding, demonstrate cost savings to providers
Integration	Coordinate the data in one registry with others, or link to other public health information systems
Performance	Need metrics to measure performance

### *Data Quality*

Data quality is a broad category encompassing many requirements. The data submitted to registries, whether electronic or paper, needs to be accurate and complete. The registry needs to ensure data security and avoid duplicate records. Immunization records must be submitted to the registry on a continual basis, and trust with providers needs to be established so that the registry can serve as the community's master record (9,13). Data submitted to registries electronically has been shown to be more complete and accurate (14). Data quality can be improved through simple steps such as more accurate data entry and bookkeeping at the practice (15).

### *Participation*

Provider participation is tied in with data quality. Having high quality data available for only a small section of the population does not serve the epidemiologic potential of these systems. Although private providers are the main source of childhood immunizations, participation has historically been lacking. Monetary costs to the practice have been cited as one of the main barriers to higher participation rates (16,17). Reducing data entry time with more intuitive user interfaces, electronic transfer of records, and improved system availability can lower costs (18). It is conceivable that the cost savings from improved efficiency and reduction in duplicate vaccinations can offset the cost of participation.

### *Economics*

Immunization registries are expensive to deploy and maintain. Sustainable funding sources, both public and private, are necessary. For its registry initiatives, the All Kids Count II project determined the participation cost was approximately \$3.91 per child, and maintaining a “nationwide network of registries” could require almost \$80 million annually (19). The Boston Immunization Information System (BIIS) required over \$345,000 in 1998 for development and maintenance of the system, with a participation cost of \$5.45 per child (20). In California, approximately \$250,000 was necessary to construct immunization registries (21).

Despite the initial startup capital, registries have the potential for savings over the long run. BIIS saved over \$26,000 in 1998 through increased efficiency for staff. The potential

cost offset from a nationwide network of registries could be over \$110 million annually (19).

### *Integration*

With a myriad of health information systems in use, a broader public health initiative would benefit from integrating immunization registries with other clinical systems. As part of the preventive medicine approach to care, children receive a variety of screening and therapeutic interventions. It is not uncommon for these data to be stored in disparate systems, which burdens providers attempting to access this information. While there are a variety of screening and therapeutic programs in practice, system integration has mainly focused on newborn dried blood spot screening, hearing screening, lead screening, immunization registries, vital registration, and the Women Infants and Children and Medicaid programs (22-25).

The Minnesota Department of Health has identified three areas to be addressed before successful integration with their immunization registries: 1) legislation for reporting, sharing, and security of data, 2) sustainable funding, and 3) addressing technical requirements. These three areas should be addressed prior to commencement of any integration project. The successful integration of childhood information systems will improve the delivery and quality of pediatric care.

## *Performance*

For an integrated registry to be useful to providers and the public, performance and progress measures are essential. Saarlal and others recommend the performance indicators used in the All Kids Count Project serve as a template for measuring the maturity of other registries (26). These indicators, developed for the All Kids Count Quantitative Indicator Survey, “[measure] timeliness of populating registries with birth and immunization data, maturity of the database, provider enrollment and participation, and immunization coverage levels” (27).

## **Current State of Immunization Registries**

According to the CDC, every state and the District of Columbia has an operating immunization registry (28). Some states, due to their size or population, have multiple registries. For example, Pennsylvania has two registries separately serving Philadelphia (Kids Immunization Database/Tracking System) and the rest of the state (Pennsylvania Statewide Immunization Information System). New York has four operating immunization registries. The idea of a single, cohesive, national registry has been studied, but due to the unique requirements of each registry and varying state laws, regional development is the most effective solution.

An annual CDC-conducted survey in 2004 showed approximately 48% of children age six or less participate in a CDC-funded registry (29). In contrast, the national goal for 2010 is to have 95% participation for children six or younger. The survey also revealed



76% of public providers and 39% of private providers submitted data to a registry during the last six months of 2004.

The national objective of 95% participation has not been met, but progress has been made. Ten of the CDC-funded initiatives (18%) have achieved the 95% participation rate for the communities they serve, and an additional seven (13%) are approaching this level. There has been a 27% increase in participation between 2000 and 2004 (30). Private provider participation must continue to increase to meet the 95% objective.

Oregon is one of the CDC identified states that has already achieved the 2010 national objective of 95% participation for children age six or younger. Between 81% and 94% of Oregon's private providers have submitted data to ALERT. All of Oregon's public providers submit data. As a result, immunization data (defined by two or more vaccination events) is available in ALERT for over 93% of pre-school age children (31).

## **About the Oregon Immunization ALERT Registry**

Before discussing the development of the epidemiologic data warehouse, it is important to understand more about ALERT. The Oregon Immunization ALERT is a statewide registry storing vaccination data for the population of children from birth through age eighteen, and recently, an increasing proportion of adult vaccination data (31). The state Department of Human Services (DHS) maintains and administers the system. Both public and private providers participate in ALERT. There are currently over 1.2 million patient records covering approximately 27 million immunizations. As with other registries, the top benefits of ALERT include complete immunization histories, official vaccination documentation for school, day care, and camp, and reminders when a child is due or past due for vaccination.

The ALERT registry has enjoyed a history of positive support from public and private practices, with 100% of public providers and over 87% of private providers submitting data since 1996. Any licensed healthcare provider in Oregon can participate in ALERT. They have the ability to not only submit data (either electronically or via bar coded forms) but also to query for data from the ALERT web site: <http://www.immalert.org>. Use of the web service is prevalent among providers; over 30,000 searches occur monthly.

## **Architecture**

Given a large user-base and vast quantity of data, an efficient infrastructure is needed to support the application. The ALERT architecture includes two servers, the registry server

and a web server, and a relational database. The registry server runs Microsoft Windows 2000 as the operating system and Microsoft SQL Server as the database application. The database has grown to approximately thirty gigabytes. The web server runs Microsoft Windows 2000, Microsoft SQL Server, and Internet Information Services.

Rather than having users of the web service access live ALERT data, which inherently carries certain risks, users access a web server that contains a weekly snapshot of ALERT. This snapshot includes demographic and vaccination information for children age birth through eighteen. Additionally, a vaccination forecast application is available that covers age birth through eight years. (A vaccination forecast contains forthcoming shots for a child, with recommended dosing dates, and the child's status.)

## Tables

The architecture of the ALERT database is built around six core data tables with numerous application-support tables. Table 3 lists the core tables and a brief description of each. A more detailed description follows. See Appendix A for a complete list of fields in the ALERT core tables.

**Table 3. ALERT core tables.**

Table	Description
Demographics	Contains the core demographics information for a child, except address, phone, and name
Phone	Stores phone number(s)
Address	Stores contact address(es)
AliasNames	Stores child's name(s)
Vaccinations	Contains the immunization events for a child
IDAlias	Serves as a deduplicated list of children in ALERT; primarily used to generate reports or for data presentation

### *Demographics*

The Demographics table contains the primary demographic information for a child. Data is stored in this table according to a single record per child per source, meaning that each provider submits data for their patients. Since children switch providers during the course of their care, this results in record duplication of the same child, because data has been submitted by multiple providers as well as billing and insurance companies. Thus, a record in the Demographics table represents only data sent by a particular provider about a particular patient. Phone number, address, and patient name are stored in separate tables. The data stored in Demographics is unlikely to change over time (for example, mother's maiden name, date of birth).

### *Phone*

The Phone table contains the current and any prior contact phone numbers for a child. This data is stored in a separate table since phone number has the potential to change over time, and ALERT users desire a historical record of phone numbers. When a provider submits data, if the phone number is different or not previously recorded, it gets stored in this table.

### *Address*

The Address table contains the current and any prior contact addresses for a child. This data is stored in a separate table since address has the potential to change over time, and

ALERT users desire a historical record of addresses. When a provider submits a record, if the address is different or not previously recorded, it gets stored in this table.

### *AliasNames*

The AliasNames table contains the child's name, and any previous or different names. This data is stored in a separate table since the name has the potential to change over time, and ALERT users desire a historical record of names. A disparity in a child's name is mainly due to submitted data inconsistency, which can result from hyphenated or multiple last names. When a provider submits a record, if the name is different or not previously recorded, it gets stored in this table.

### *Vaccinations*

The Vaccinations table stores immunization records for a child. Data is stored in this table according to a single record per child per source. As with Demographics, this results in duplication of vaccination events being recorded. For example, the healthcare provider, a billing service for that provider, or an insurance company can all report the same vaccination event. Confounding this is the problem of a single vaccination antigen being coded differently as a result of different billing or clinical coding nomenclatures. Therefore, de-duplication of vaccination data is necessary when tracking vaccination history.

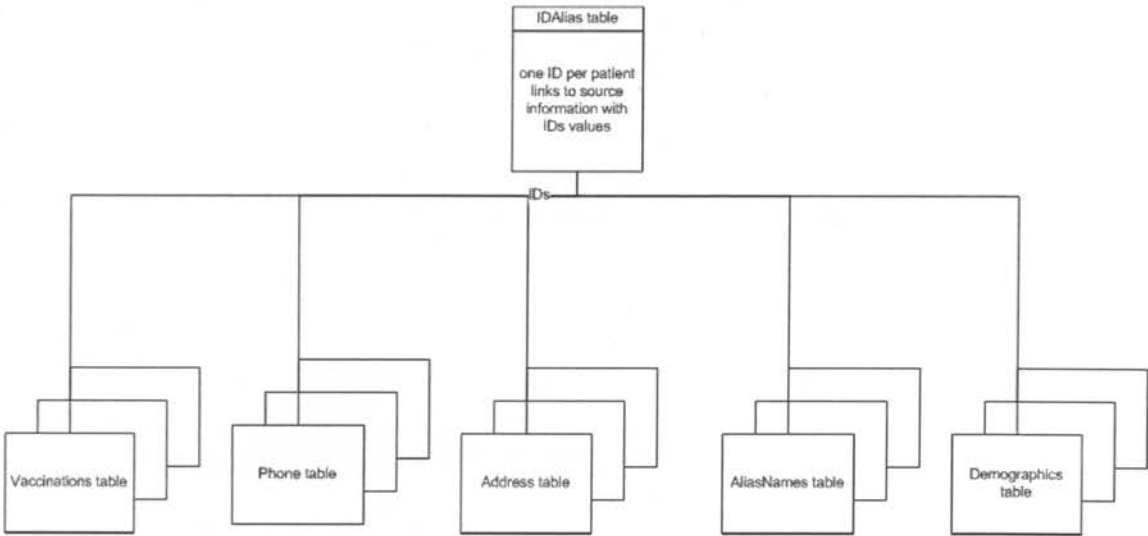
### *IDAlias*

The IDAlias table serves as the de-duplicated list of children in Demographics. This table is relied upon when data needs to be presented to the user, such as when creating reports. De-duplication is accomplished by matching key fields (for example, mother's maiden name and date of birth) to reach a level of confidence that multiple records belong to the same child.

### **Table Relationships**

The data field 'IDs' serves as the linking key between the ALERT core tables. It is the primary key in IDAlias and the foreign key in the Demographics, Phone, Address, AliasNames, and Vaccinations. 'IDs' is defined as a one-to-one relationship between IDAlias and Demographics, and a one-to-many relationship between IDAlias and Phone, Address, AliasNames, and Vaccinations. Figure 1 presents a high level overview of the ALERT core tables.

**Figure 1. ALERT core tables.<sup>2</sup>**

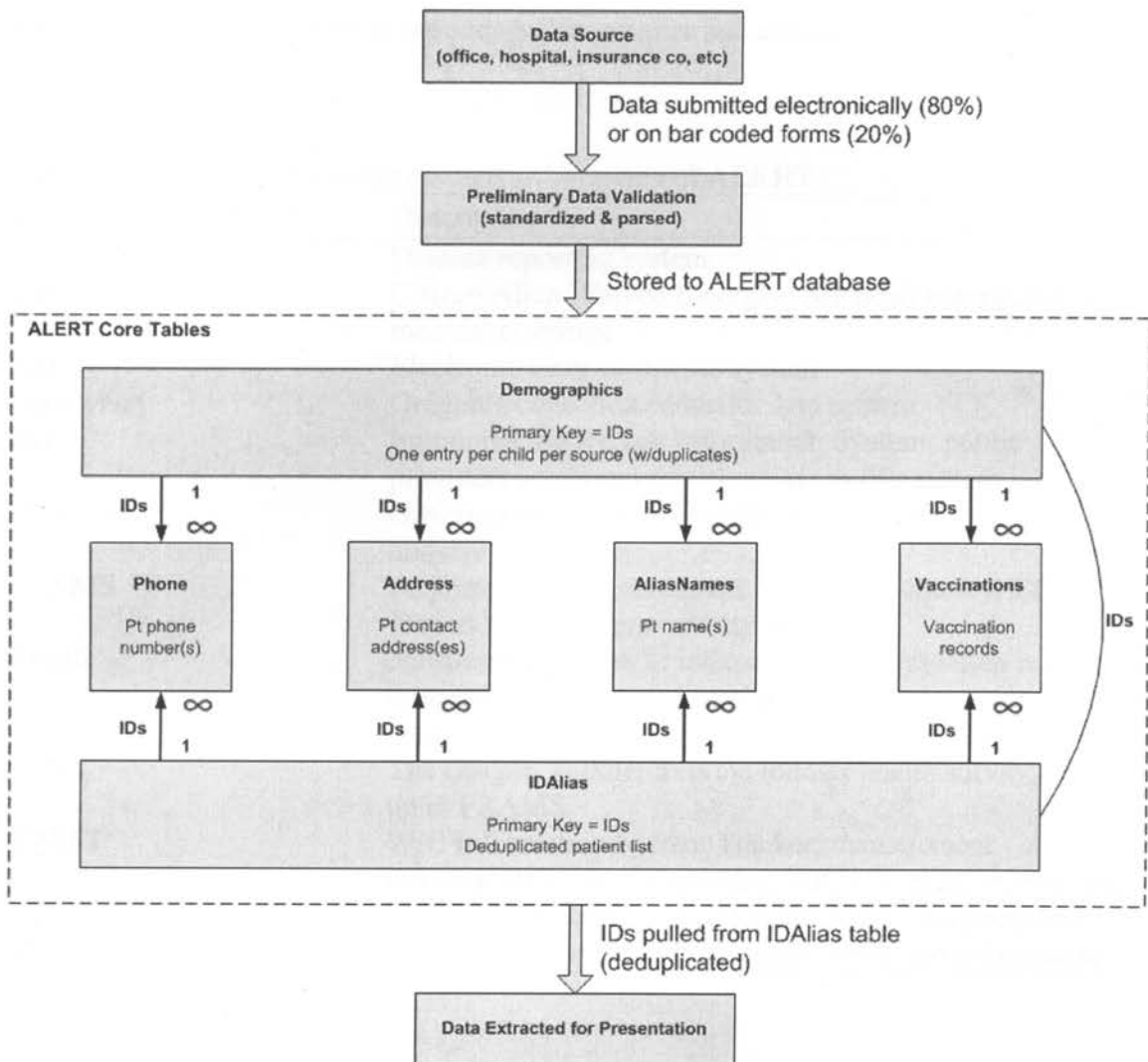


**Workflow**

The ALERT registry workflow, outlined in Figure 2, entails data submission to database storage. The process begins with a vaccination event being submitted to ALERT. This occurs electronically (80%) or via faxed or mailed bar coded forms (20%). Data submitted electronically is standardized and checked for simple errors. Bar coded forms are manually keyed into the system. After data entry, the data is parsed and placed into the correct tables in ALERT. Providers then have the option of receiving reports of submitted data to check for validity, which unfortunately only occurs a minority of the time. More often, data is stored into ALERT without any additional data validation.

<sup>2</sup> Courtesy of Donald M. Dumont/Oregon Department of Human Services

**Figure 2. ALERT registry workflow.**



## Secondary Data & Subsystems

There are a variety of internal and external data sets that are incorporated with ALERT, and are required for public health and epidemiologic research. Some of the more commonly used data sets include Vital Records, IRIS, and Medicaid. ALERT also has subsystems that provide added functionality to the registry. Recall is the most notable of these subsystems. The Recall system provides clinical decision support; that is, it



generates post card reminders for parents and providers for due or past due vaccinations.

Table 4 summarizes the major secondary data sources and subsystems.

**Table 4. Secondary data sources and subsystems of ALERT.**

Data Set	Description
CASES	Disease reporting system
CAWEM	Citizen Alien Waived Emergent Medical; emergency medical coverage
EBC	Electronic birth certificate system
FamilyNet	Oregon's consolidated health data system
IRIS	Immunization Record Information System; public providers send immunization data to this system
Medicaid	Government sponsored health insurance for underprivileged
PRAMS	Pregnancy Risk Assessment Monitoring Survey; CDC-funded survey of postpartum women
Recall	Notification system to indicate when vaccination is past due; generates postcard reminders for due/past due vaccinations
TOTS	The Oregon Toddler Survey; toddler health survey; follow up to PRAMS
TWIST	WIC Information System Tracker; management information system supporting WIC (Women, Infants, and Children; government supplemental nutrition program)
VR	Vital Records system; birth/death certificate information

### **Motivation for a Research Data Warehouse**

The data stored in ALERT is not in an optimal format for research projects. The current ALERT database is optimized as a clinical system including the use of algorithms for instance identification (i.e., demographic de-identification) and for entity recognition (i.e., vaccination de-identification). These may not be the same algorithms that are required for a research database. DHS staff has indicated it is challenging to obtain data from ALERT for research purposes. Therefore, DHS would like to create from the clinical database a second database that can be used for surveillance or research. Oregon Health & Science University has been chartered to propose a model and develop the data warehouse.

## **Methodology**

In order to recommend an appropriate strategy for the development of the epidemiologic data warehouse, it is necessary to ascertain how the system will be used. To accomplish this, sample research questions were obtained from DHS and a questionnaire was distributed to DHS researchers that surveyed how the data warehouse would be used in their research.

### **Sample Research Questions**

DHS research interests varied broadly with respect to the secondary data sets required. Consequently, one of the requirements of the data warehouse will be to incorporate the various data sets (Table 4) into the research database, or minimally, provide the linking fields to the external data. The following are examples of current DHS research interests:<sup>3</sup>

- Basic evaluation of ALERT recall efforts plus an evaluation of recall effects on providers practices. Basic evaluation is to compare impact of recall process for 28-35 month olds across providers who actively participate (review recall report and update data prior to postcards) and those providers who do not participate (no recall report review). Second evaluation is to look at a provider's cohorts of kids to see if participation in the 24 month recall influences the provider's early (2, 4, 6 months) immunization and reporting patterns.

---

<sup>3</sup> Courtesy of Martha P. Skiles/Oregon Department of Human Services

- Examine factors associated with under-immunization by 24 months (for combined 4:3:1 or 4:3:1:3:3<sup>4</sup>, and a few single antigens, e.g. PCV7<sup>5</sup>). Factors to include recall postcard data (e.g., dates of pre-recall to provider, dates of postcards, and labels of which antigens recalled), linked to birth certificate info to obtain child/parent/prenatal service info. Then examine if factors differ by region.
- From the birth file data we can identify type of birth attendant (e.g., doctor, midwife, nurse midwife, etc.). DHS would like to look at type of birth attendant as a factor associated with up-to-date rate and associated with getting immunization data in ALERT.
- Basic evaluation of timing of uptake for new vaccines and/or new recommendations. Example is when DHS compared a new accelerated DTaP schedule for 3 counties for 6 months. DHS needed to look at different age cohorts (eligible for shots before, during, and after recommendations) from both intervention and non-intervention counties, to determine if the accelerated schedule intervention was actually implemented in the 3 counties.
- DHS often confuse or can not tell the difference between 1) factors that are associated with lack of immunization with 2) factors that are associated with under-reporting or poor capture of immunizations. For example, if a set of factors in the birth record are associated with a higher probability of moving out of state at some point, those factors will also, spuriously, be associated with lower immunization UTD rates in ALERT. How do we use our data to identify items

---

<sup>4</sup> The 4:3:1:3:3 immunization series consist of four DTaP (Diphtheria, Tetanus, acellular Pertussis), three IPV (Inactivated Poliovirus Vaccine), one MMR (Measles, Mumps, Rubella), three Hib (*Haemophilus influenza* type b), and three Hep B (Hepatitis B).

<sup>5</sup> Pneumococcal Conjugate Vaccine

that are associated predominantly with non-capture of data, so that we can use these factors as stratifiers or controls against any proposed factors associated with under-immunization?

- Are kids, and at some point adults, getting more shots than they need because clients do not remember or do not have good (accurate) records, so providers repeat shots? To do this DHS would need good consolidated client records with appropriate de-duplication of shots based on close reporting/administration dates.

## **Questionnaire**

To ensure the epidemiologic data warehouse accurately reflects the needs of the researchers, we submitted a questionnaire to DHS. This questionnaire was divided into two parts: 1) identification of research needs, and 2) identification of data warehouse requirements. After the questionnaires were returned, the results were summarized and distributed to the participants. A follow-up meeting took place with DHS to discuss the results and reach a consensus for each survey item. The full questionnaire and summary results can be found in Appendices B and C.

We made a few fundamental assumptions about the data warehouse, which helped to focus the questionnaire on particular items of interest. The first assumption was that the research data warehouse would be a static snapshot from the live ALERT database. When possible, it is preferable to work with data that is derived from the primary source, so as not to burden the live system with queries and potentially corrupt the original data. The next assumption was the ALERT database contained superfluous data that was not

necessary for a research system. This would allow us to consolidate the data down to the minimally required fields, improving performance and decreasing size requirements. Our last assumption was the duplicate children and vaccinations present in ALERT could be consolidated to a single entry per child and per immunization event.

The questionnaire was distributed to seventeen members of the ALERT staff at DHS. Ten surveys were returned in the allotted time. Active ALERT users submitted eight of these surveys; prospective users submitted the other two. Participants included epidemiologists, research analysts, immunization coordinators, and ALERT management.

The first part of the survey included five questions focused on the types of research carried out with ALERT. We learned the following:

- ALERT is used for population and individual (patient-specific) studies. Therefore, identifiers to individual children are necessary in the data warehouse.
- Age ranges for study cohorts varied, but generally included birth through school age. We reached a consensus at the follow-up meeting to initially track through the age eight.
- Study cohorts are tracked over time, from year to year. The data warehouse will consequently be built upon a series of cohort databases, partitioned by birth year.
- Researchers depend on access to a multitude of internal and external data sets.

These are summarized in Table 4.

- Within each of the secondary data sets, researchers use specific information that, in some cases, is unique for each research study. As a result, we need to minimally provide the linking fields for each data set.

The second part of the survey focused on specific requirements for the data warehouse.

We learned the following:

- The data warehouse needs to be updated from the live ALERT database at least annually. Some researchers expressed an interest in having a semiannual update. At the follow-up meeting, a consensus was reached for an annual update to coincide when cleaned data is received from the Vital Records system.
- Researchers who use ALERT for individual (patient-specific) studies require the ability to identify a particular child from their cohort. Therefore, individual child identifiers are necessary in the warehouse.
- Currently ALERT stores records in the Demographics table as one record per child per source. Researchers agreed it would be acceptable to consolidate to a single record per child, essentially merging multiple providers for a particular child into a single record.
- No preference was expressed for a vaccination de-duplication algorithm. We agreed at the follow-up meeting that immunization data should not be overly processed. An antigen specific algorithm could be developed that uses a window of valid days (determined for each shot antigen) to identify if multiple records refer to the same immunization event.

- When consolidating from a single record per child per source to a single record per child, we will need to keep the entire history of a child with respect to provider, county of residence, and other fields in Demographics that may change over time.



## **Data Model**

The data model for the warehouse outlines the overall architecture (databases, tables, and fields) necessary for DHS staff to efficiently and effectively carry out research. Although the data model represents a compromise among the various research interests of staff, and therefore may not include all the required data for a particular study, the entire ALERT data is still accessible to researchers. By including the core identifiers for ALERT, as well as the secondary data sources, information not included in the warehouse can still be accessed. The tradeoff is the additional overhead required to manually retrieve information from these sources. Our goal is to include in the warehouse the most commonly used ALERT fields and the linking fields for the secondary data sets.

## **Databases**

The data warehouse will be stratified by birth cohort databases. Each birth year exists in a separate database with the prior eight years comprising the active cohorts. For example, the 2006 data warehouse will have the following cohort databases: 2005, 2004, 2003, 2002, 2001, 2000, 1999, and 1998. The 1998 database contains all children born in 1998, and vaccination data through 2005 (eight years). Similarly, the 1999 database will contain seven years vaccination data, the 2000 database contains six years, and so on. The data warehouse will be turned (i.e., updated from ALERT) annually, every April or May, to coincide with updated and cleaned Vital Records data for the prior calendar year. To accommodate ongoing studies during the annual turn, the old databases will be archived for one to two years. Each birth database will have an identical structure. Table 5 illustrates the databases and birth cohorts for a sample data warehouse.

**Table 5. Databases of birth cohorts.**

<i>2006 Data Warehouse</i>								
<b>Birth Year Database</b>	<b>2005</b>	<b>2004</b>	<b>2003</b>	<b>2002</b>	<b>2001</b>	<b>2000</b>	<b>1999</b>	<b>1998</b>
Includes vaccination data for birth year and subsequent years enumerated below.		2005	2004	2003	2002	2001	2000	1999
			2005	2004	2003	2002	2001	2000
				2005	2004	2003	2002	2001
					2005	2004	2003	2002
						2005	2004	2003
							2005	2004
								2005

Since each birth database will have an identical structure, the remainder of the discussion of the data model focuses on an individual database within the warehouse. The tables, fields, and relationships apply to all eight birth cohort databases.

### Tables

Four core tables are envisioned in a birth cohort database: Demographics, Vaccinations, TemporalData, and IDsLookUp. Table 6 presents the core tables and a brief description of each, with a more detailed description following.

**Table 6. Birth cohort database core tables.**

Table	Description
Demographics	Contains the minimum required demographic information for each child, one record per child
Vaccinations	Contains the vaccination records, one record per immunization event
TemporalData	Contains current and historical data for fields that change over time, such as county and provider
IDsLookUp	Resolves 'IDs' value(s) for a particular child in Demographics

### *Demographics*

This table is analogous to the Demographics table in ALERT. The data will be consolidated from a single entry per child per source to one entry per child. Several of the fields in the ALERT Demographics table (such as the administrative data – see Appendix A) are unnecessary in a research database, and therefore will not be imported.

Demographics will also contain a subset of the Vital Records data, which the researchers identified as the most frequently used data set in ALERT. The primary key for this table will be a new identifier, ‘DemoID’. This field will be used for linking to the other tables in the database.

### *Vaccinations*

The Vaccinations table is also analogous to the Vaccinations table in ALERT. The data will be de-duplicated to a single record per vaccination event based on a previously agreed upon antigen algorithm. The primary key will be a new identifier, ‘VaxID’. The linking key for this table will be the ‘IDs’ field imported from ALERT.

### *TemporalData*

TemporalData is a new table specific to the birth cohort database. In ALERT, data that changes over time (i.e., temporal data) is stored in the tables Phone, Address, and AliasNames. Since we will not be importing these tables (to minimize the size of the database), the necessary data fields from these tables will be placed in the TemporalData table. Furthermore, when we consolidate from a single record per child per source to a single record per child, we will need to store disparate data from the multiple records.

TemporalData will be used to store current and historical data for fields that change over time, such as county and provider. Each record will be date stamped and sorted reverse chronologically, most recent entries appearing first for a child. The primary key will be a new identifier, 'RefID'. The linking key for this table will be the 'IDs' field imported from ALERT.

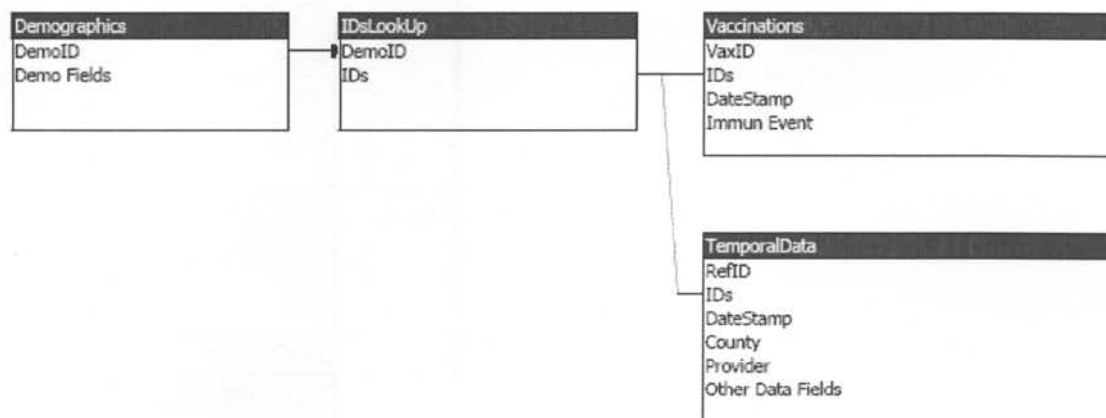
### *IDsLookup*

ALERT stores records using a unique identifier known as 'IDs'. This identifier is distinct for a child/source combination. All of the core tables in ALERT have this identifier for relational linking. Since we are consolidating from one record per child per source to one record per child (and therefore consolidating multiple 'IDs' to a single record), we will need a method of resolving the 'IDs' from a single child demographic record. An IDsLookup table will serve this purpose by mapping a single, unique ID ('DemoID') from Demographics to multiple 'IDs' values for the supporting tables (i.e., Vaccinations and TemporalData). Conversely, the IDsLookup lookup table provides a method for resolving back to Demographics from supporting tables by identifying the unique child ID ('DemoID') for each 'IDs' mapped to the same child. 'IDs' will be the primary key since each value is unique in this table.

### **Table Relationships**

To aid in the visualization of the structure and interaction between the tables in the birth cohort database, Figures 3 through 6 are depictions of a sample birth cohort database built using Microsoft Access.

**Figure 3. Relationships between the birth cohort database tables.**



Demographics has a one-to-many relationship to IDsLookUp, via the DemoID primary key (unique to each entry/child). As discussed previously, when consolidating from one record per child per source to one record per child, multiple 'IDs' values can occur for each child. IDsLookUp in turn has a one-to-many relationship to Vaccinations and TemporalData, via the existing 'IDs' key in ALERT that will be imported for each record.

**Figure 4. Demographics linked to IDsLookUp.**

		DemolD	Demo Fields
	-	1001	Child1
		IDs	
	+	9001	
	+	9002	
	+	9003	
	*	0	
	-	1002	Child2
		IDs	
	+	9005	
	*	0	
	-	1003	Child3
		IDs	
	+	9010	
	+	9011	
	+	9012	
	*	0	
▶		0	

Figure 4 displays three sample demographic records for Child1, Child2, and Child3. The 'DemoID' field is the unique, primary key for the Demographics table. Shown below each record are the corresponding 'IDs' values (from IDsLookUp) that were encountered when merging from one record per child per source to one record per child. Again, these are necessary to be tracked in order to link to the supporting tables (Vaccinations and TemporalData).

**Figure 5. Demographics linked to Vaccinations.**

		DemolD		Demo Fields	
▶	-	1001		Child1	
		IDs			
▶	-	9001			
		VaxID	DateStamp	Immun Event	
▶		1	4/29/2000	DTaP	
		2	4/29/2000	Hib	
		3	4/29/2000	HepB	
	*	(AutoNumber)			
	-	9002			
		VaxID	DateStamp	Immun Event	
		4	5/2/2001	MMR	
	*	(AutoNumber)			
	-	9003			
		VaxID	DateStamp	Immun Event	
		5	6/16/2005	MMR	
		6	6/16/2005	DTaP	
	*	(AutoNumber)			
	*	0			
-	-	1002		Child2	
		IDs			
	-	9005			
		VaxID	DateStamp	Immun Event	
		7	8/12/2000	DTaP	
		8	8/12/2000	IPV	
		9	8/12/2000	Hib	
		10	8/12/2000	HepB	
	*	(AutoNumber)			
	*	0			
+	-	1003		Child3	
*		0			

Figure 5 shows the three sample demographic records (Child1, Child2, Child3) and the corresponding immunization events (as linked through IDsLookup to Vaccinations). Note that each 'IDs' can, and most likely always will, contain multiple vaccination records.

**Figure 6. Demographics linked to TemporalData.**

DemolD		Demo Fields				
-	1001	Child1				
		IDs				
	-	9001				
		RefID	DateStamp	County	Provider	Other Data Fields
		1	10/20/2000	Multnomah	Smith JM	
	*	(AutoNumber)				
	-	9002				
		RefID	DateStamp	County	Provider	Other Data Fields
		2	10/31/2000	Multnomah	Kaiser	
	*	(AutoNumber)				
	-	9003				
		RefID	DateStamp	County	Provider	Other Data Fields
		3	5/1/2004	Multnomah	Dobson VA	
	*	(AutoNumber)				
*		0				
-	1002	Child2				
		IDs				
	-	9005				
		RefID	DateStamp	County	Provider	Other Data Fields
		4	1/31/2000	Grant	Medicaid	
	*	(AutoNumber)				
*		0				
-	1003	Child3				
		IDs				
	-	9010				
		RefID	DateStamp	County	Provider	Other Data Fields
		5	6/14/2000	Columbia	Blue Cross	
	*	(AutoNumber)				
	+	9011				
	+	9012				
*		0				
▶	0					

Figure 6 illustrates the three sample demographic records (Child1, Child2, Child3) with the corresponding temporal data records (as linked through IDsLookUp to TemporalData). This is the intended method of preserving data that changes over time. Although each record in TemporalData in this example contains both a county and provider, only one field will be required, as these fields are independent. The additional



temporal data fields (in this figure shown as “Other Data Fields”) will be determined post-import. The intent is to show specific fields that contain data that change over time.

## **Extraction, Transformation, and Loading**

Extraction, transformation, and loading (ETL) entail the steps of data acquisition to the warehouse. Extraction involves selecting the appropriate data from ALERT and other data sets, transformation conditions the data for input, and loading is the act of importing into the corresponding tables in the warehouse (32). A brief overview is presented, however, specific ETL methodology is outside the scope of this project.

The majority of data imported to the warehouse from ALERT will not require much transformation. Most of the fields will be a direct extract from ALERT, and load to the warehouse. During the record consolidation phase, certain fields will be moved from Demographics to TemporalData. In addition, fields in the Phone, Address, and AliasNames tables will be moved to TemporalData.

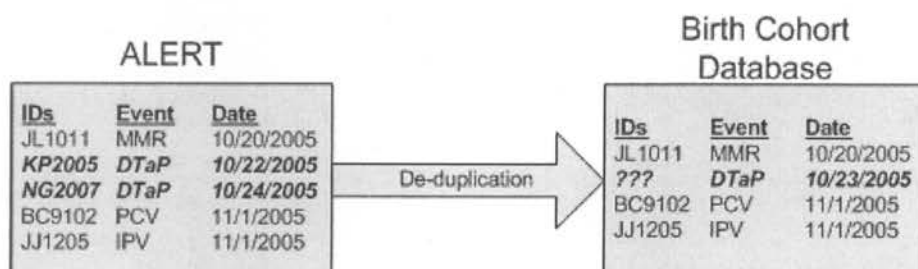
The values for the IDsLookUp table will be derived from the Demographics table during the consolidation phase. For example, if three records belong to the same child, the three 'IDs' values will be entered into IDsLookUp as three distinct entries, where each entry carries the same 'DemoID' identifier. Figure 4 displays one possible scenario of 'IDs' mappings to 'DemoID'.

De-duplication will occur for the Demographics and Vaccinations tables in ALERT. Both de-duplication efforts will rely on algorithms chosen by DHS. We recommend the algorithms minimally process the data, and thus, err on the side of too many (duplicates) rather than too few (missing) records.

Vaccination de-duplication is a special case that requires additional processing to resolve the necessary 'IDs' values. For example, if two immunizations are recorded in ALERT with the administration dates of the 20<sup>th</sup> and 24<sup>th</sup> of the month, and it is determined that these separate records indicate the same vaccination, the de-duplication algorithm will consolidate these records. Since each original record (i.e., the 20<sup>th</sup> and 24<sup>th</sup>) can also carry a unique 'IDs' value (if each event was submitted by a different source, such as an insurance company and doctor's office), this raises the issue of which 'IDs' value to record for the de-duplicated record.

There are two solutions to address this special case. The first is to arbitrarily choose one of the 'IDs' values from the original record, and use that value for the de-duplicated record. The second is to assign a new, unique 'IDs' value to the de-duplicated record. Provided that the entry in the IDsLookUp table correctly maps the 'IDs' value to 'DemoID', either solution is acceptable. In either case, the ability to map from the de-duplicated record back to the original records in ALERT is compromised. Figure 7 shows a vaccination de-duplication scenario.

**Figure 7. Vaccination de-duplication.**



## Maintenance

The research data warehouse is a rolling snapshot of ALERT. It will be turned every April or May, to coincide with updated and cleaned Vital Records data for the prior calendar year. The oldest birth cohort database will be archived in the new warehouse for one to two years, to accommodate in-progress studies. For example, the 2007 data warehouse will contain current immunization record databases from 1999 through 2006, and an archive database for 1998, assuming only one prior year is kept. Since the archive database is not updated in the annual turn of the warehouse (but rather just copied over from the 2006 to 2007 warehouse), it will only have vaccination records through 2005. Table 7 depicts this visually.

**Table 7. Databases of birth cohorts with archived year.**

<i>2007 Data Warehouse</i>									
<b>Birth Year Database</b>	<b>2006</b>	<b>2005</b>	<b>2004</b>	<b>2003</b>	<b>2002</b>	<b>2001</b>	<b>2000</b>	<b>1999</b>	<b>1998</b>
<i>1998 is the archive year</i>		2006	2005	2004	2003	2002	2001	2000	1999
			2006	2005	2004	2003	2002	2001	2000
Includes vaccination data for birth year and subsequent years enumerated below.				2006	2005	2004	2003	2002	2001
					2006	2005	2004	2003	2002
						2006	2005	2004	2003
							2006	2005	2004
								2006	2005

## **Summary and Conclusion**

The ALERT epidemiologic data warehouse is a powerful tool that will enable researchers to readily access immunization data. The data warehouse is a rolling snapshot of the production ALERT immunization registry, and will be updated annually. Data will be imported from ALERT, Vital Records, Recall, and various other secondary data sources necessary for DHS research studies. The warehouse will be built upon a series of birth cohort databases, drawn from a subset of children (that is, birth through age eight) in ALERT. Records in each database will be consolidated to a single record per demographics entry (one record per child) and a single record per vaccination event.

Four tables will contain the core data in the warehouse: Demographics, Vaccinations, TemporalData, and IDsLookUp. The Demographics table holds the pertinent information for each child. Vaccinations contains the entire vaccination history for the birth cohort. TemporalData keeps track of information for each child that has the potential to change over time (for example, provider or county of residence), and therefore serves as a historical record. IDsLookUp provides a mapping from the single record per child model of the warehouse to the single record per child per source model of ALERT.

Although the data model for this project is specific to the ALERT architecture, the methodology used to develop it is applicable to other immunization registries. The questionnaire was successful in soliciting the proper feedback from DHS, as well as fostering appropriate dialogue among the research staff at the follow up meeting. The

framework presented here serves as an appropriate starting point for the development of an effective and powerful immunization research system.

## References

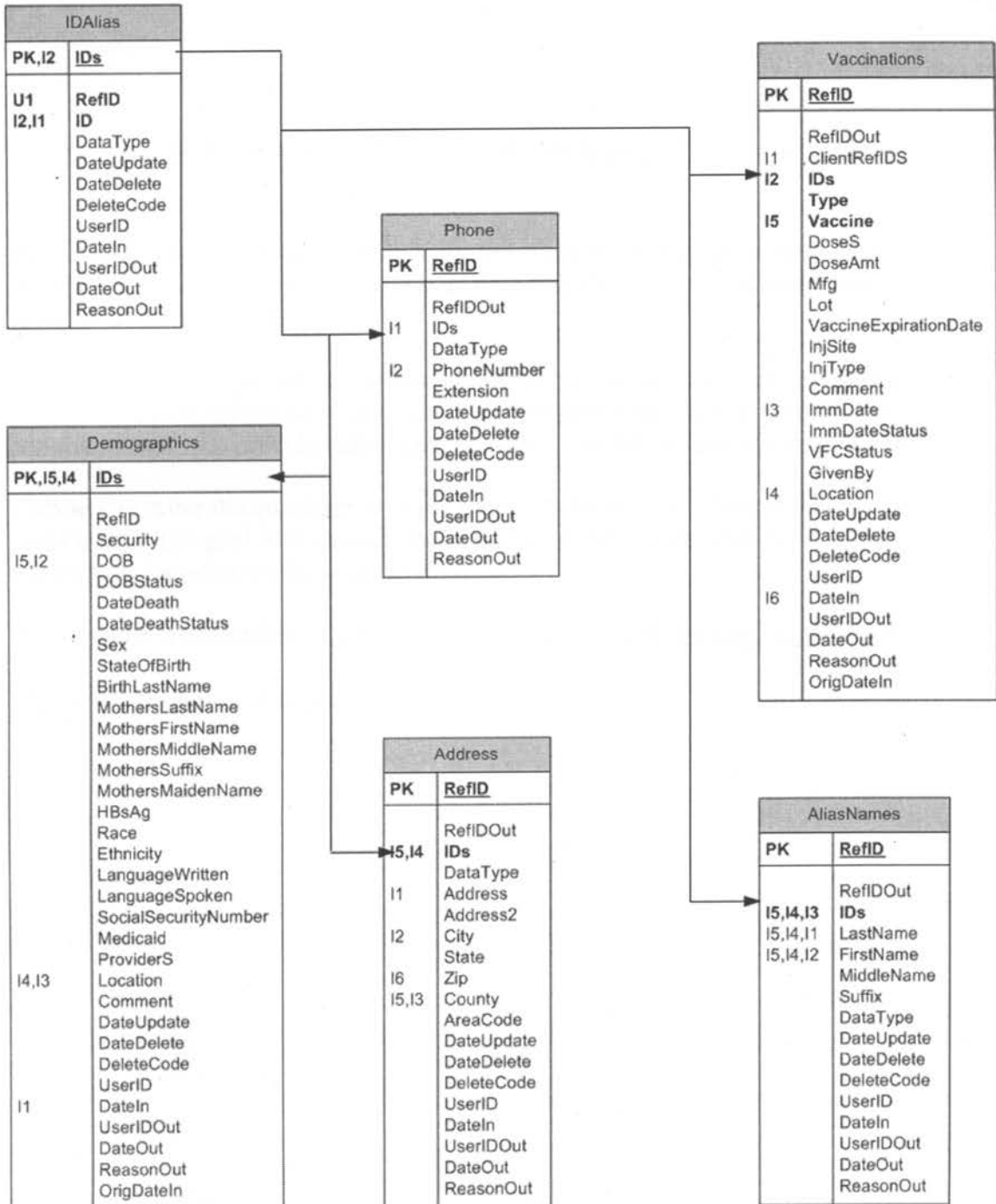
1. Centers for Disease Control and Prevention (CDC). Initiative on Immunization Registries. MMWR Recomm Rep. 2001 Oct 5;50(RR-17):1-17.
2. Centers for Disease Control and Prevention (CDC). Development of community- and state-based immunization registries. CDC response to a report from the National Vaccine Advisory Committee. MMWR Recomm Rep. 2001 Oct 5;50(RR-17):1-17.
3. Healthy People 2010 - Objective 14-26. Available at: <http://www.healthypeople.gov/document/html/objectives/14-26.htm>. Accessed May 13, 2006.
4. All Kids Count - Grant Results. 2002; Available at: <http://www.rwjf.org/portfolios/resources/grantsreport.jsp?filename=akc.htm&iaid=141&gsa=1>. Accessed May 13, 2006.
5. Centers for Disease Control and Prevention (CDC). General Recommendations on Immunization: Recommendations of the Advisory Committee on Immunization Practices (ACIP) and the American Academy of Family Physicians (AAFP). MMWR Recomm Rep. 2002 Feb 8;51(RR-2):1-36.
6. Freeman VA, DeFries GH. The challenge and potential of childhood immunization registries. Annu Rev Public Health 2003;24:227-246.
7. Centers for Disease Control and Prevention (CDC). Recommended childhood and adolescent immunization schedule--US, 2006. Ann Pharmacother. 2006 Feb;40(2):369-371.
8. Kempe A, Beaty BL, Steiner JF, Pearson KA, Lowery NE, Daley MF, et al. The regional immunization registry as a public health tool for improving clinical practice and guiding immunization delivery policy. Am J Public Health 2004 Jun;94(6):967-972.
9. Kempe A, Steiner JF, Renfrew BL, Lowery E, Haas K, Berman S. How much does a regional immunization registry increase documented immunization rates at primary care sites in rural colorado? Ambul Pediatr. 2001 Jul-Aug;1(4):213-216.
10. Linkins RW. Immunization registries: progress and challenges in reaching the 2010 national objective. J Public Health Manag Pract. 2001 Nov;7(6):67-74.
11. All Kids Count. Immunization Registries Save more than they Cost. 2000; Available at: <http://www.allkidscount.org/iz/whatsnew/cost.html>. Accessed May 13, 2006.

12. Tierney CD, Yusuf H, McMahon SR, Rusinak D, O'Brien MA, Massoudi MS, et al. Adoption of reminder and recall messages for immunizations by pediatricians and public health clinics. *Pediatrics* 2003 Nov;112(5):1076-1082.
13. Davidson AJ, Melinkovich P, Beaty BL, Chandramouli V, Hambidge SJ, Phibbs SL, et al. Immunization registry accuracy: improvement with progressive clinical application. *Am J Prev Med.* 2003 Apr;24(3):276-280.
14. Kolasa MS, Chilkatowsky AP, Clarke KR, Lutz JP. How complete are immunization registries? The Philadelphia story. *Ambul Pediatr.* 2006 Jan-Feb;6(1):21-24.
15. Samuels RC, Appel L, Reddy SI, Tilson RS. Improving accuracy in a computerized immunization registry. *Ambul Pediatr.* 2002 May-Jun;2(3):187-192.
16. Clark SJ, Cowan AE, Bartlett DL. Private provider participation in statewide immunization registries. *BMC Public Health* 2006 Feb 15;6(1):33.
17. Glazner JE, Beaty BL, Pearson KA, Elaine Lowery N, Berman S. Using an immunization registry: effect on practice costs and time. *Ambul.Pediatr.* 2004 Jan-Feb;4(1):34-40.
18. Rask KJ, Wells KJ, Kohler SA, Rust CT, Cangialose CB. The cost to providers of participating in an immunization registry. *Am J Prev Med.* 2000 Aug;19(2):99-103.
19. Horne PR, Saarlal KN, Hinman AR. Costs of immunization registries: experiences from the All Kids Count II Projects. *Am J Prev Med.* 2000 Aug;19(2):94-98.
20. McKenna VB, Sager A, Gunn JE, Tormey P, Barry MA. Immunization registries: costs and savings. *Public Health Rep.* 2002 Jul-Aug;117(4):386-392.
21. Fontanesi JM, Flesher DS,Jr, De Guire M, Lieberthal A, Holcomb K. The cost of doing business: cost structure of electronic immunization registries. *Health Serv Res.* 2002 Oct;37(5):1291-1307.
22. Hinman AR, Atkinson D, Diehn TN, Eichwald J, Heberer J, Hoyle T, et al. Principles and core functions of integrated child health information systems. *J Public Health Manag Pract.* 2004 Nov;Suppl:S52-6.
23. Fehrenbach SN, Kelly JC, Vu C. Integration of child health information systems: current state and local health department efforts. *J Public Health Manag Pract.* 2004 Nov;Suppl:S30-5.



24. Hoyle T, Swanson R. Assessing what child health information systems should be integrated: the Michigan experience. *J Public Health Manag Pract.* 2004 Nov;Suppl:S66-71.
25. Papadouka V, Schaeffer P, Metroka A, Borthwick A, Tehranifar P, Leighton J, et al. Integrating the New York citywide immunization registry and the childhood blood lead registry. *J Public Health Manag Pract.* 2004 Nov;Suppl:S72-80.
26. Saarlal KN, Edwards K, Wild E, Richmond P. Developing performance measures for immunization registries. *J Public Health Manag Pract.* 2003 Jan-Feb;9(1):47-57.
27. All Kids Count Quantitative Indicator Survey. 2001; Available at: <http://www.allkidscount.org/iz/researchandeval/survey.htm>. Accessed Jun 10, 2006.
28. Immunization Information Systems - State IIS Staff. 2006; Available at: <http://www.cdc.gov/nip/registry/contacts/contact-prog-tech.htm>. Accessed May 18, 2006.
29. Centers for Disease Control and Prevention (CDC). Immunization information system progress--United States, 2004. *MMWR Morb Mortal Wkly Rep.* 2005 Nov 18;54(45):1156-1157.
30. Centers for Disease Control and Prevention (CDC). Progress in development of immunization registries--United States, 2000. *MMWR Morb Mortal Wkly Rep.* 2001 Jan 12;50(1):3-7.
31. Canavan BC. Immunization Registries: Data Use for Population Assessment.
32. Khan AH. Data Warehousing 101: Concepts and Implementation. Khan Consulting and Publishing; 2003. p25.

## Appendix A. ALERT Core Tables with Fields.<sup>6</sup>



<sup>6</sup> Courtesy of Donald M. Dumont/Oregon Department of Human Services

## **Appendix B. ALERT Research Data Warehouse Questionnaire.**

### **ALERT Research Data Warehouse Questionnaire**

*Purpose:* To identify researcher needs for the creation of an epidemiological data warehouse. The data warehouse will contain a snapshot of essential data from ALERT to assist in research studies.

*Directions:* This questionnaire is divided into two sections: 1) Identification of research needs, and 2) Identification of data warehouse requirements. Space for additional comments has been provided after each question – please include any pertinent feedback.

*Follow Up:* After the questionnaires are returned, the results will be shared with the participants. The goal is to identify the best strategy for the development of the data warehouse, focusing on the researcher needs.

Please return the completed questionnaire to Neal Goldstein ([goldsten@ohsu.edu](mailto:goldsten@ohsu.edu)).

Thank you for your participation.



### Identification of Research Needs

**1. Does your research use ALERT for population or individual (patient specific) level studies?**

\_\_\_\_ Population      \_\_\_\_ Individual      \_\_\_\_ Both

**2. What are the patient age(s) in the study cohorts that you most commonly use?**

**3. Do you need to follow patient cohorts over time? That is, will you track the same group of patients from year to year?**

\_\_\_\_ Yes      \_\_\_\_ No

**If Yes, please provide an idea of the time interval typically looked at:**

**4. What internal/external data sets are necessary for your research in ALERT?**

\_\_\_\_ Vital Records      \_\_\_\_ IRIS      \_\_\_\_ CASES      \_\_\_\_ Medicaid

\_\_\_\_ Recall

**Please indicate any other data sets that are not listed above:**

**5. From the data set(s) you identified above, specifically what information is being used (e.g., birth certificate number, birth attendant, recall postcard data, etc.)?**

## Identification of Data Warehouse Requirements

**6. How often should the research data warehouse be updated from ALERT? That is, how often should the data snapshot occur?**

☐ Semiannually    ☐ Annually    ☐ Other, please specify:

**7. Will you need to be able to identify a specific patient from your cohort? That is, should there be the ability to de-anonymize patients?**

☐ Yes    ☐ No

**8. Would it be acceptable to consolidate the data to one demographic record per child (instead of one record per child per source) to minimize the size of the database?**

☐ Yes    ☐ No

If No, please explain:

**9. Is there a preferred vaccination deduplication algorithm?**

☐ Yes    ☐ No

If Yes, please identify:

**10. When merging records, there will be instances where information changes for a child over time (e.g., county of residence, provider, etc.). Is there a preferred method for resolving fields that have disparate info?**

☐ Most Recent    ☐ Random    ☐ Weighting  
☐ Keep All (i.e. don't resolve)    ☐ Other, please specify:

**11. Please provide any additional comments or questions relating to the development of the epidemiological data warehouse.**

Appendix C. Summary of Questionnaire Results.



**ALERT Research Data Warehouse Questionnaire  
Survey Results**

Identification of Research Needs

**1. Does your research use ALERT for population or individual (patient specific) level studies?**

☐ Population      ☐ Individual      ☐ Both

*Results*

7 identified Both, 1 identified Individual, 1 identified Population, 1 N/A

**2. What are the patient age(s) in the study cohorts that you most commonly use?**

*Results*

Ranges included All ages; 0 – 6,7,8 yrs; 12 – 26,35 mos

**3. Do you need to follow patient cohorts over time? That is, will you track the same group of patients from year to year?**

☐ Yes      ☐ No

**If Yes, please provide an idea of the time interval typically looked at:**

*Results*

Ranges included 1 yr, 1 – 3 yrs, 0 – 6 yrs, and longer/shorter intervals

**4. What internal/external data sets are necessary for your research in ALERT?**

\_\_\_\_ Vital Records    \_\_\_\_ IRIS    \_\_\_\_ CASES    \_\_\_\_ Medicaid  
\_\_\_\_ Recall

**Please indicate any other data sets that are not listed above:**

*Results*

8 identified VR, 5 identified IRIS, 4 identified CASES, 5 identified Medicaid, 4 identified Recall

Additional data sets included: reportable disease data (e.g. Hep B info), PRAMS, TOTS, census data measures (e.g., % poverty in community), TWIST (WIC)

**5. From the data set(s) you identified above, specifically what information is being used (e.g., birth certificate number, birth attendant, recall postcard data, etc.)?**

*Results*

Typical demographics (especially county), birth certificate items (especially #), sex, race/ethnicity, place of birth, SES, vax status, anything have risk factors on (?), birth attendant, payer at birth, public vs private clinic used, clinic location

*From CASES:* name, dob, diagnosis, disease info

*From EBC:* cert #, race/ethn, address, name, dob, HBsAg status of mom, screening & birth dose data

*From FamilyNet:* name, id#, dob, client status (active/inactive), provider

*From Medicaid:* enrollment category (especially CAWEM), diagnostic categories (e.g., birth defects), visits (e.g., well child care)

*From Recall:* reminder/recall postcard elements

*From VR:* demographics, outcome

Identification of Data Warehouse Requirements

**6. How often should the research data warehouse be updated from ALERT? That is, how often should the data snapshot occur?**

\_\_\_\_ Semiannually    \_\_\_\_ Annually    \_\_\_\_ Other, please specify:

*Results*

6 identified Annually, 4 identified Semiannually

**7. Will you need to be able to identify a specific patient from your cohort? That is, should there be the ability to de-anonymize patients?**

\_\_\_\_ Yes

\_\_\_\_ No

*Results*

6 identified Yes (need to de-anonymize), 4 identified No

**8. Would it be acceptable to consolidate the data to one demographic record per child (instead of one record per child per source) to minimize the size of the database?**

\_\_\_\_ Yes

\_\_\_\_ No

**If No, please explain:**

*Results*

8 identified Yes (one record/pt), 1 Maybe (with agreed upon system for accurately deciding which source to use), 1 N/A

**9. Is there a preferred vaccination deduplication algorithm?**

\_\_\_\_ Yes

\_\_\_\_ No

**If Yes, please identify:**

*Results*

1 identified Yes (minimal that doesn't overly process shot date records, could discuss algorithm for each antigen), 4 identified No, 5 Unsure

**10. When merging records, there will be instances where information changes for a child over time (e.g., county of residence, provider, etc.). Is there a preferred method for resolving fields that have disparate info?**

\_\_\_\_ Most Recent

\_\_\_\_ Random

\_\_\_\_ Weighting

\_\_\_\_ Keep All (i.e. don't resolve)

\_\_\_\_ Other, please specify:

*Results*

7 identified Keep All (don't resolve), 2 identified Most Recent, 1 Unsure



**11. Please provide any additional comments or questions relating to the development of the epidemiological data warehouse.**

*Results*

Flexibility in data use is very important

Security of data is critical

Would be useful to have calculated variables for each antigen – both total and valid shots