



VA Corporate Data Warehouse (CDW)

A VIREC Resource Guide

VA Information Resource Center (VIREC)

June 2012

VIReC Resource Guide: VA Corporate Data Warehouse

Overview

Introduction This Resource Guide is designed to help health services researchers and other users of the Corporate Data Warehouse (CDW) understand CDW origins, structure, content and utility for VA research.

The Guide is a product of the VA Information Resource Center (VIReC), a national resource center of the U.S. Department of Veterans Affairs (VA) Health Services Research and Development Service.

Support The VA Information Resource Center (VIReC) is supported by Health Services Research and Development Service grant SDR 98-004.

Suggested citation VA Information Resource Center. VIReC Resource Guide: VA Corporate Data Warehouse. Hines, IL: U.S. Dept. of Veterans Affairs, Health Services Research and Development Service, VA Information Resource Center, Jun. 2012.

Contents This document contains the following topics:

Topic	See Page
CDW Description	3
Data Currently Available	5
Research Utility- Data Quality	9
Research Utility- Strengths	10
Research Utility- Limitations	11
CDW Documentation	12
Additional Resources	13

CDW Description

Introduction

The U.S. Department of Veterans Affairs Corporate Data Warehouse (CDW) is a national data repository comprising data from several Veterans Health Administration (VHA) clinical and administrative systems.

The CDW operates within the U.S. Department of Veterans Affairs (VA) Office of Information & Technology's Field Operations Business Intelligence Service Line.

Objective

The objective of the CDW is to incorporate data from multiple data sets throughout the VHA into one standard database structure to facilitate reporting and data analysis at the enterprise level.

The goal of CDW's staff is to provide data and tools to support management decision making, performance measurement, and research objectives.

CDW SharePoint® site

The CDW staff maintains a SharePoint® site where they provide information and support for users. The CDW SharePoint site is available only on the VA Intranet [1].

Source data systems

VistA, the electronic medical record system utilized across the VA, is the primary source of data for the CDW. Additional sources of CDW data include:

- VA Regional Data Warehouses
- VHA Decision Support System (DSS)
- VHA National Patient Care Database (NPCD)
- VA Compensation and Pension Exams
- MyHealthVet, the VA's personal health record system

The CDW is still in development and more data sources will be added in the future.

Location

The CDW is located at the Austin Information Technology Center in Austin, Texas. Support personnel are located at several sites across the country.

Continued on next page

CDW Description, Continued

Design

CDW data are stored in a relational database made up of a collection of data domains. Data tables belonging to each domain can be merged at the level of the individual patient or at various organizational levels as desired.

This database structure can simplify data management for studies that require data from more than one source data system.

For example, researchers can request VistA inpatient, outpatient, and pharmacy data and DSS data all merged into one file. Furthermore, while the Decision Support Office originally created a separate National Data Extract (NDE) file for each year and each Veterans Integrated Services Network (VISN), CDW staff compressed those files into a single table for each of the NDE types (Lab, Lab Results, Pharmacy, etc.) Each of the tables includes all VISNs and all of the years 2005 through 2009.

Data update frequency

Data within the CDW are updated based on a predetermined schedule. This schedule is available from the CDW SharePoint site. Follow the steps below to access the update schedule.

Step	Action
1	Visit the CDW SharePoint site, available on the VA Intranet
2	On the “CDW Home” page, click on the picture of a table labeled “CDW Domain Priority & Status”
3	View the table of update frequencies by data type

Replaced data values are not saved

When the CDW is updated, changed data values are written over, not maintained in a historical record. These incremental updates are, however, flagged using the following business rules.

When data are:	Then the following business rules apply:
Updated	Changed values are overwritten and flagged with a “Last Update” date
Deleted	Data that has been deleted from a source database is retained in CDW and flagged as “deleted” (Delete flag =Y)

Continued on next page

CDW Description, Continued

Case identifiers In CDW data, cases are identified by multiple types of identifiers which allow analysts to identify unique individuals and to link cases within CDW data and with data obtained elsewhere.

Case Identifier	Definition
Scrambled Social Security Number	A VA-created identifier generated by applying an algorithm to an enrollee's Social Security Number (SSN)
PatientSID	Patient Surrogate ID (SID) number, a CDW-assigned identifier Note: PatientSID is specific to facility so if a patient is seen in more than one facility, he or she will have more than one PatientSID.
PatientIEN	Patient Internal Entry Number (IEN) is an individual's identification number in his/her local VistA system
PatientICN	Patient Integration Control Number (ICN) is the VHA's unique patient identification number generated by the Master Veteran Index.
Social Security Number	Unique personal identifier issued by the Social Security Administration. Note: Investigators may apply for permission to access data identified by SSN only if their investigation requires such access.

Data Currently Available

Introduction The CDW's relational database is comprised of data domains. Each domain is a set of tables with a common theme; usually the theme indicates the application in the VistA electronic health record system from which most of the data elements come (e.g., Vital Signs or Mental Health Assessment).

Domain types Data domains are categorized as Production or Raw.

Production status domains CDW Production domains comprise data that have been organized to support flexible querying among tables and views. Production domains are divided into two tiers. The table below provides the definition of the tiers.

Tier	Definition
1	The set of views available for querying on the main Production databases. The only transformations applied to the source data are structural, specifically, reorganization to facilitate more flexible querying. No filtering of records, editing of content or business rules are applied. Occasionally some structures are added, for example, to map a set of codes to a standard set.
2	A set of database objects (views, tables, stored procedures, etc.) developed for a specific business context such as a performance measure or measures, a disease cohort, a provider's panel of patients, or a research project. Typically these data sets are drawn from a subset of the Tier 1 data and then complex logic is frequently applied to generate a data mart. Mapping a set of codes to a standard set is more common within a Tier 2 than within a Tier 1 domain.

Raw status domains A direct extract from the source system, reflecting the source system structure, not modeled, standardized, or indexed. Sometimes a Raw domain is a work area for a new data domain in preparation for Production status. In addition, Decision Support System (DSS) National Data Extracts and other conversions of SAS files to SQL tables are categorized as Raw domains.

Continued on next page

Data Currently Available, Continued

Available domains

The CDW staff adds domains as they are developed so it is recommended that researchers check the CDW SharePoint site for the most up-to-date information on data domain availability. The table below shows how to find out what data domains are currently available.

Step	Action
1	Visit the CDW SharePoint site, available on the VA Intranet
2	On the “CDW Home” page, click on the picture of a table labeled “Domain Priority & Status”
3	View the table of update frequencies by data type

As of April 30, 2011, the following CDW Production domains are available and refreshed with new data nightly:

- Appointments
- Consults
- CPRS Orders
- Encounters
- Health Factors
- Immunizations
- Inpatient
- Inpatient Census
- Laboratory (Chemistry)
- Mental Health Assessment
- Outpatient
- Patient (Demographics, etc.)
- Primary Care Management Module
- Pharmacy, Bar Code Medication Administration
- Pharmacy, Outpatient
- Staff (Demographics, Provider Type, etc.)
- Vital Signs

Dates available Most CDW data are available for records dated October 1, 1999 and forward.

Types of data that were created after 1999 are available from the data’s creation date forward.

For example, the earliest Mental Health Assessment data were created in and are available in the CDW from approximately October 2007.

Access

Introduction

Research access to CDW data is granted only to Institutional Review Board (IRB) approved projects and is managed through the DART (Data Access Request Tracker) system.

Access approval process

The table below provides a general outline of the process for obtaining approval for research access to CDW data.

Step	Action						
1	Submit your data request using DART Note: DART is available on the VA Intranet [2].						
2	Administrative review: National Data Systems (NDS) reviews your data request first to make sure all documentation is complete and appropriate						
3	Data Steward’s review: NDS sends your data request to appropriate offices for approval <table border="1" data-bbox="566 978 1401 1253"> <thead> <tr> <th data-bbox="566 978 774 1052">Data Steward</th> <th data-bbox="774 978 1401 1052">Approval Offices</th> </tr> </thead> <tbody> <tr> <td data-bbox="566 1052 774 1131">NDS</td> <td data-bbox="774 1052 1401 1131"> <ul style="list-style-type: none"> • VHA Information Access & Privacy Office • VHA Health Care Security Office. </td> </tr> <tr> <td data-bbox="566 1131 774 1253">Not NDS</td> <td data-bbox="774 1131 1401 1253"> <ul style="list-style-type: none"> • VHA Information Access & Privacy Office • VHA Health Care Security Office • Office of the data steward </td> </tr> </tbody> </table>	Data Steward	Approval Offices	NDS	<ul style="list-style-type: none"> • VHA Information Access & Privacy Office • VHA Health Care Security Office. 	Not NDS	<ul style="list-style-type: none"> • VHA Information Access & Privacy Office • VHA Health Care Security Office • Office of the data steward
Data Steward	Approval Offices						
NDS	<ul style="list-style-type: none"> • VHA Information Access & Privacy Office • VHA Health Care Security Office. 						
Not NDS	<ul style="list-style-type: none"> • VHA Information Access & Privacy Office • VHA Health Care Security Office • Office of the data steward 						
4	Privacy and Security review for requests for real SSNs: If you are requesting real SSNs, NDS will send your application to the VHA Office of Research and Development for review Note: Scrambled SSN requests do not need this approval.						
5	Final approval: NDS awards final approval to your application and alerts VINCI staff tasked with data extractions.						

Continued on next page

Access, Continued

Hints for successful applications

Prior to completing the DART application for CDW data, researchers are encouraged to visit the CDW site on the VA Intranet [1]. This site contains information on data available from the CDW and the CDW's terminology for data which researchers will want to use on their DART applications.

Research data requests to the CDW must be consistent with the study protocol and IRB approval.

Researchers should communicate to the CDW their file format preference (text or SQL format) and what file documentation they would like from the CDW (e.g., data layouts for text data, data element formats).

Access process

Researchers do not work directly on the CDW server. Instead researchers are provided by VINCI personnel with customized data extracts. The table below depicts steps to follow to gain data access, once access privileges have been granted.

Step	Action
1	VINCI data manager assigned to create your data extract contacts you by email.
2	Data manager works with you to create your customized data extract Note: You may send either a detailed cohort description or case identifiers to the CDW so CDW staff can produce the desired extract file.
3	Data manager transfers the extract to your workspace on the VINCI server or to another server within the VA firewall.

Using VINCI as an analysis environment

Use of VINCI for CDW data analysis is not required but it offers several advantages described on the VINCI Central web site on the VA Intranet [3].

For example, VINCI provides access to statistical and natural language processing software applications, abundant server space and, most importantly, a secure environment on the VA Intranet for analysis of patient data.

Continued on next page

Research Utility- Data Quality

Introduction

CDW personnel employ quality assurance protocols intended to safeguard the quality of CDW data. In addition, the VHA Office of Informatics and Analytics (OIA) Data Quality Program plans to explore the quality of data within the CDW.

CDW quality monitoring

CDW Extract, Transform and Load (ETL) routines transfer data from the source databases to the CDW. ETL routines also monitor data transfers by tracking percent deviation from previous transfer counts, process timing, and error log counts, all in order to ensure that the CDW contains an accurate replica of data as they exist at their source. The table below shows examples of rules included in the ETL routines and the basic data quality principles they check.

Rule	Data Quality Principle Checked
Procedural rule	Makes sure all rows are accounted for
Data quality rule	<ul style="list-style-type: none">• Confirms number of rows is within 1 standard deviation of previous counts• Confirms number of rows having a particular field null does not exceed a prescribed threshold

Note: CDW ETL procedures are also designed to identify rare but clear and problematic violations of data integrity, for example, a date in a field that should contain a “Yes” or “No”. CDW’s ETL process deletes these values and substitutes a value of ‘E’ for error.

Continued on next page

Access, Continued

VHA Data Quality group's CDW validation

The VHA Office of Informatics and Analytics (OIA) Data Quality Program plans to explore the quality of data within the CDW and to provide data quality guidance to the CDW as necessary. Their plan includes examining CDW metadata to determine if the requirements that CDW was asked to meet have been met and also looking for problems such as truncations of fields.

In a 2010 NDS comparison of CDW data with the VHA's Medical SAS[®] data sets, a record-to-record comparison for 25% of the records for each fiscal year in each of those sources found a greater than 99% agreement in field value for the great majority of data fields compared.

CDW personnel are in frequent communication with the staffs of both NDS and OIA, and several adjustments have already been made to increase the accuracy of the CDW.

Guidance for researchers

As with any new data source, researchers will want to explore their data, examining whether data fields are populated completely and with expected values, and consider the implications of their findings for their research analyses.

Research Utility- Strengths

Introduction Strengths of the CDW include the capture of data not currently available in any other national data source, the number and variety of its contributing data sources, and the frequency of data updates. The relational database design allows more flexibility for database architecture than do some other database management formats.

Data available only from the CDW CDW is the only national VA data source for data on Vital Signs, Health Factors and text notes. Lab (Chemistry) results and Radiology results are available for all tests from 1999 and forward. These data are critical components of clinical and health services research.

Decision Support System National Data Extract (NDE) files that have been compressed into a single table for each of the NDE types (lab, lab results, pharmacy, etc.) for all VISNs and all years 2005 through 2009 are available only from the CDW.

Merged data from multiple sources available Data from several sources can be requested in one data request and the requested data can often be merged into one file of person-level data, a time saver for both investigators and data managers. This feature will become more advantageous as the CDW's data holdings expand.

Frequency of updates For some studies current data is critical. CDW Production data are updated nightly. Raw data are generally updated much less frequently: weekly, bi-monthly or even less frequently. Update frequency for each CDW data domain is posted on the CDW SharePoint site on the VA Intranet.

Relational database advantage While MedSAS data sets allow a fixed, limited number of codes to be recorded for a particular data element, CDW's relational data are not so restricted. For example, the MedSAS Outpatient file includes a maximum of ten ICD-9-CM procedure codes per encounter. Because of its relational database design the CDW Outpatient Encounters data can include all ICD codes present in a veteran's VistA files.

Research Utility- Limitations

Introduction Limitations of the CDW include its limited selection of available data, unknown data quality for many data elements, lack of stable reference files and the requirement that analysts become familiar with new data storage conventions.

Additional data sources needed At this time, the CDW's selection of data is growing but limited and so, for most research questions, CDW users must also obtain data from additional sources.

Data quality CDW data quality is largely untested. To begin with, the CDW's data quality is greatly dependent on the quality of VistA data which is also largely untested. Furthermore, few head to head comparisons of VistA and CDW data have been accomplished to date. Moreover, while out-of-range values are cleaned from MedSAS data, any errors and out-of-range values in VistA data will be found in CDW data.

No stable reference file CDW is a constantly updated warehouse holding no stable reference files comparable to the MedSAS data set final, fiscal year files. Research needs to be replicable by future researchers in order to be deemed valuable and the lack of reference files for future researchers to access may be problematic.

Continued on next page

Research Utility- Limitations, Continued

Flags required Researchers new to CDW data will want to ask for and use data elements called flags created by the CDW data architects. These flags often indicate something important about the data. Flags applicable to each CDW data domain are listed on the domain’s metadata report on the CDW SharePoint site on the VA Intranet. The table below shows a few of the available flags.

Flag	Identifies
Delete	Records that were deleted as errors from a source database are not deleted from the CDW but are instead flagged as “deleted” (Delete Flag = Y).
TestPatientFlag	Data entered while testing computer systems
EnteredInErrorFlag	Vital Sign Data records that entered in error

**Probable
Unlabeled
Test Patient
Flag**

Some probable test patients in the CDW are not flagged with the “TestPatientFlag”. The Data Quality Analysis (DQA) team within the OIA Data Quality Program is developing a *CDWPossibleTestPatientFlag* to identify both labeled and unlabeled probable test patients.

In early 2012, the DQA team observed cases of likely test patient records in the CDW Production database. These cases have unlikely values for identifiers such as patient name and SSN, but their identifier values do not meet the guidelines for identifying test patients specified by VHA Directive 2006-036, *Data Quality Requirements for Identity Management and Master Patient Index Functions* [4].

The DQA team worked with CDW data architects to develop a method to identify probable unlabeled test patients. DQA’s Guidance Report, *“Identifying Probable Unlabeled Test Patients”* contains a description of the method, decision rules, and sample source code for implementation of the decision rules. The Guidance Report is available on the CDW SharePoint site on the VA Intranet [5].

CDW Documentation

Introduction CDW documentation is available on the CDW SharePoint site on the VA Intranet [1].

Most CDW data originates in VistA. A useful source for VistA documentation is the VistA Metadata Repository of the VHA Data Architecture Repository (DAR), available on the VA Intranet [6].

CDW Metadata Reports Select the Metadata tab on the CDW SharePoint site to see Metadata Reports for each domain. The table below describes some of the contents of the Metadata Reports.

CDW Metadata Report Contents	Description
Links to domain sub-models (also called Entity Relationship (ER) diagrams)	Click on a domain sub-model name to see: <ul style="list-style-type: none">• A diagram of the tables belonging to the domain• Some text about the dates for which the domain's data are available• The source files for the domain's tables, usually the VistA file and field numbers
Links to tables	Click on a table name to see: <ul style="list-style-type: none">• Field Names• Field types and formats• Source file and field• Field descriptions• Primary and foreign keys

Note: Field descriptions that are available within the VistA system are extracted to the CDW Metadata Reports column headed "VistA Field Description". Unfortunately, VistA field descriptions tend to be cryptic and/or technical. CDW staff provides descriptions for some fields in the column headed "Field Description".

Continued on next page

CDW Documentation, Continued

Domain Action Team documents

The Data Stewardship Program, a division of the OIA Data Quality Program, is working with the CDW to recruit subject matter experts for Domain Action Teams for future CDW domains. The teams will determine which VistA data elements should be included in the CDW data for their domains. To expedite this work, the CDW's SharePoint site now has web pages on which team members can share documents.

Note: As each domain is finalized, the domain team's web page will be opened up to all CDW SharePoint users to promote greater understanding of the data and to share sequel queries others have developed.

Data Architecture Repository

CDW requirements for VistA-based domains will be defined based on DAR metadata so metadata for all the elements in VistA-based CDW domains should appear in the DAR. The VistA Metadata Repository has a convenient Search function, for finding metadata by data element name, VistA File Name or File number.

Additional Resources

Introduction Archives of past training sessions, answers to frequently asked questions and more general education about the CDW can be found on the CDW site. Details and additional information sources are listed below.

CDW User Support Frequently asked questions and their answers can be found on the Support tab of the CDW site, along with links to archived training materials.

The CDW now offers General Announcement and Training Announcement alerts by email. To sign up, select Announcements in the left banner on the CDW site Support page. Click on the Subscribe button for either or both alert types.

Austin Help Desk The Austin Information Technology Center's National Service Desk (NSD) offers technical support for users of the CDW 24 hours of every day.

NSD Contact Method	Contact Information
Help Desk Phone	512-326-6780 or 888-326-6780
TDD	512-326-6638
Email	nsd-aus@va.gov

VINCI VINCI users can get support through the VINCI Service Desk from 8:30am to 4:30pm Mountain Time, Monday through Friday.

VINCI Contact Method	Contact Information
Service Desk Phone	801-588-5212
Email	VINCI@va.gov

Continued on next page

Additional Resources, Continued

VIREC

The VA Information Resource Center (VIREC) is a good resource for VA researchers with questions about CDW data.

VIREC Contact Method	Contact Information
Help Desk Email	virec@va.gov
General website	http://www.virec.research.va.gov
Help Desk Phone	708-202-2413
Intranet website	http://vaww.virec.research.va.gov
Physical Address	VA Information Resource Center (151V) Health Services Research and Development Service Department of Veterans Affairs Edward Hines, Jr. VA Hospital 5000 South 5 th Avenue Hines, IL 60141

Works Cited

1. U.S. Department of Veterans Affairs, Office of Information and Technology, Business Intelligence Service Line. "CDW Home." Accessed April 16, 2012. (See Appendix A for VA Intranet URL.)
 2. U.S. Department of Veterans Affairs, Office of Informatics and Analytics, Health Information Governance, National Data Systems. "DART Process for Standard NDS Request." Accessed April 16, 2012. (See Appendix A for VA Intranet URL.)
 3. U.S. Department of Veterans Affairs, Office of Information Technology and Veterans Health Administration Office of Research and Development. "VINCI Central". Accessed April 16, 2012. (See Appendix A for VA Intranet URL.)
 4. U. S. Department of Veterans Affairs, Veterans Health Administration. *Data Quality Requirements for Identity Management and Master Patient Index Functions: VHA Directive 2006-036*. Washington, D.C.: U.S. Dept. of Veterans Affairs, Veterans Health Administration, June 1, 2006. Rescinded or Expiration Date: 06/30/11. Accessed Apr. 16, 2012. http://www1.va.gov/vhapublications/ViewPublication.asp?pub_ID=1434
 5. U.S. Department of Veterans Affairs, Office of Information and Technology, Business Intelligence Service Line. Mar. 16, 2012. *Data Quality Analysis Guidance Report - Identifying Probable Unlabeled Test Patients*. Accessed Apr. 16, 2012. (See Appendix A for VA Intranet URL.)
 6. U.S. Department of Veterans Affairs, Office of Information and Technology, Product Development Division. "*VHA Data Architecture Repository: Vista Metadata Repository*." Accessed Apr. 16, 2012. (See Appendix A for VA Intranet URL.)
-

Appendix A

The Intranet URLs for these references are available from the VA Intranet version of this guide or through the VIREC Help Desk (virec@va.gov).

Table A1. VA Intranet Web sites referenced in this Guide.

Ref #	Date Accessed	Name of Reference
1	April 16, 2012	“CDW Home” Web page
2	April 16, 2012	“DART Process for Standard NDS Request.” Web page
3	April 16, 2012	“VINCI Central” Web page
5	April 16, 2012	<i>Data Quality Analysis Guidance Report - Identifying Probable Unlabeled Test Patients</i>
6	April 16, 2012	“VHA Data Architecture Repository: Vista Metadata Repository” Web page
