# THE VIRTUAL DATA WAREHOUSE (VDW) AND HOW TO USE IT

## Table of Contents



Multicenter research can be greatly facilitated by the Health Care Systems Research Network Virtual Data Warehouse (VDW). This section provides context and tools to help investigators discover the potential of the VDW and navigate obstacles to using it in research projects.

## Overview

One of the hallmarks of the HCSRN is the variety and amount of electronic administrative data about the health and health care utilization of the sites' enrollees and the adoption of electronic medical records at most of the sites. To make these data more easily accessible for multicenter research projects, we have constructed a virtual data warehouse—the VDW. The VDW is not a centralized data warehouse—it is "virtual", consisting of parallel databases set up identically at each of the HCSRN sites (see Figure 1) that can be easily merged across sites. These databases have been constructed by extracting data directly from the local electronic data systems and reconfiguring them to use standard variable names and coded values. Because the data elements are pulled directly from a wide variety of working data systems at the individual sites, they may include different coding patterns and other site-specific idiosyncrasies. Therefore, the VDW is not an analytic dataset. It is a very rich collection of information from which analytic datasets can be constructed.

With the VDW, much of the preparatory work for pooling existing data across multiple sites has been done in advance. A project's analyst writes a program based on the VDW data dictionary. The program is then sent to the other participating sites to be run locally with the output files securely transferred to the project. With input from the site investigators and programmers, merged analytic datasets are then constructed.

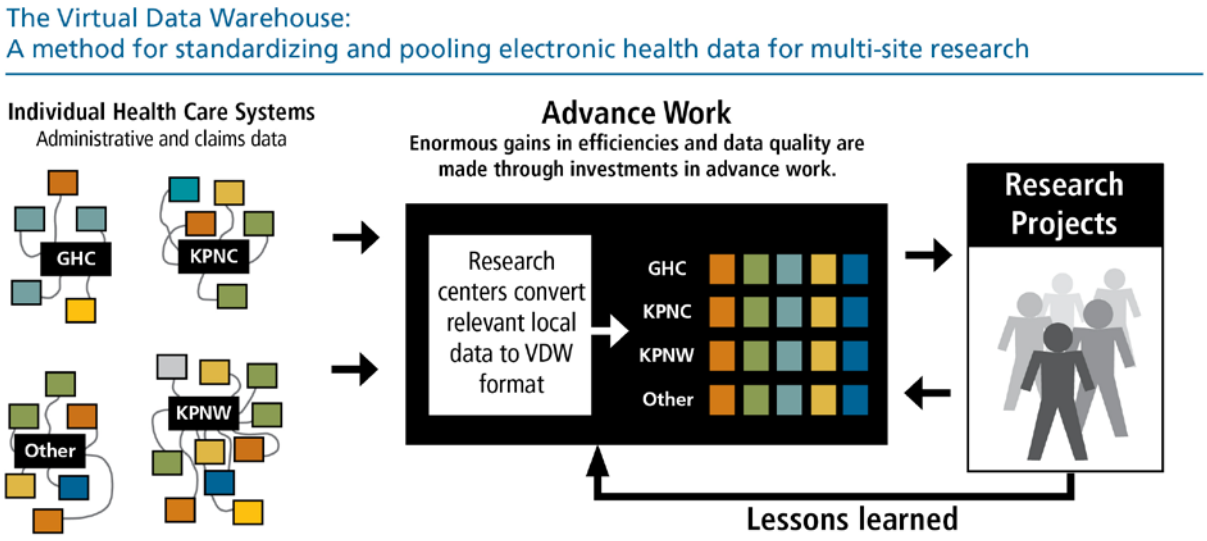
The VDW is constantly being expanded and improved as each project uses it. It contains data that support a wide variety of research studies, including studies using surveys and/or chart abstraction, as well as those directly aimed at analysis of electronic administrative data. For

example, the VDW has been used to identify subjects for extensive chart abstraction in a study of treatment and outcomes for older women with early stage breast cancer. A current project to develop a test of oral health literacy used the VDW to select potential subjects residing in neighborhoods with a variety of average educational levels. Another project developed a project-specific add-on to the VDW to assemble standardized data extracted from electronic text of pap smear reports. Projects analyzing cancer treatment have identified the variety of sources containing data on the receipt of chemotherapy by cancer patients across a number of the sites. The VDW is a work in progress with constant on-going development. The cross-site team of programmers has frequent conference calls and in-person meetings about VDW improvements.

## Example VDW projects

A great deal of research can be conducted using only VDW data. Commonly VDW data supplements other data gathered from subjects, as well – such as survey data, data or specimens from research clinic visits, and so on. A list of example projects is provided on our website.

## Figure 1. The HCSRN VDW and how it works



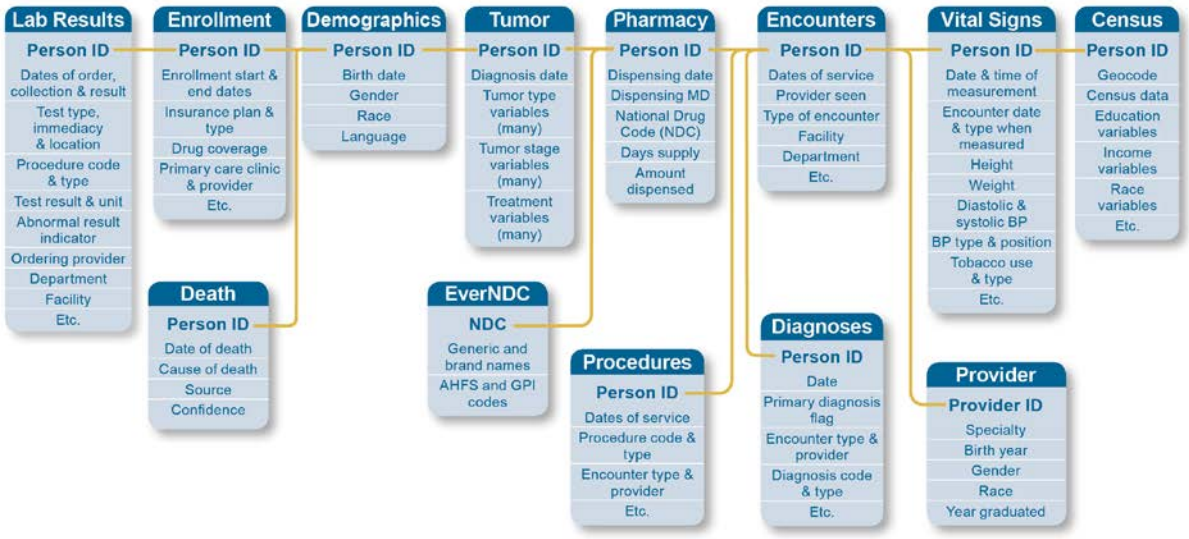

Standardized VDW data sets are already well established at most HCSRN sites. Site programmers and data managers are actively involved in on-going discussions and quality control efforts and are familiar with the process of using the VDW for research projects.

## Data Areas

As of March 2013, VDW data domains include:

- **Demographics** contains date of birth, gender, race and ethnicity.
- **Enrollment** is based on health plan membership enrollment with indicators of insurance types, benefits, and effective dates of coverage.
- **Encounters** characterizes outpatient visits and inpatient stays, including the associated diagnosis and procedure codes, type of encounter, provider seen, facility and discharge disposition.
- **Procedures** consists of all performed procedures including evaluation and management, surgery, laboratory, radiology, and immunization. Currently only performed procedures are captured and include various procedure coding systems (CPT-4, HCPCS, ICD-9-CM, insurance claims Revenue Codes).
- **Diagnoses** includes dates, diagnosis codes and codes types, primary diagnosis flag and diagnosing provider.
- **Providers** includes information on the providers such as specialty, age, gender, race and year graduated.
- **Cancer/Tumor Registry** is based on the Surveillance, Epidemiology and End Results (SEER) program standards as many HCSRN sites are SEER sites. The domain consists of detailed stage and grade, date of diagnosis, dates of treatment initiation, and is by far the most complex domain of the VDW.
- **Pharmacy Dispensing** consists of pharmacy dispensing and claims and includes date of dispensing, National Drug Code or GPI code (to standardize across sites), therapeutic class, days supply, and amount dispensed. These data are widely used to assess pharmacy-based disease and co-morbidity classification systems.
- **Vital Signs** are collected at most in-person encounters and include height, weight and blood pressure readings. Tobacco use and type is also included.
- **Laboratory Values** was originally limited to hemoglobin A1C, serum creatinine, international normalized ratio, fasting blood glucose, and serum potassium. Sites are adding laboratory values to this table through a timed priority list with 57 types of lab tests included or in process at this time.

## Figure 2: HCSRN VDW data domains




Detailed data dictionaries for each VDW file are available elsewhere on our website.

## Catalog of VDW Macros

Site programmers have developed a catalog of SAS macros, or reusable code. VDW users can employ the macros to define standard methods for sampling and data abstraction within the VDW.

Commonly used VDW SAS macros include:

- Flexible definitions of "continuous enrollment"
- First disenrollment after an index date
- Comorbidity scores including Charlson and RxRisk
- Age calculation
- Cases with invasive breast cancer between specific dates
- Outpatient pharmacy fills for a given sample
- Outpatient pharmacy fills for a given list of national drug codes (NDCs)
- Counts of fills for a given list of NDCs
- BMI calculation
- Vital sign measures for a given sample
- Census data for a given sample

## Steps for Using the VDW

The VDW is ideally suited for supporting the development of grant proposals, carrying out preliminary studies, identifying subjects, and assembling study data.

## Using the VDW for grant proposal development

- Discuss the potential study with an investigator at your site experienced with HCSRN multicenter projects (e.g., CRN, MHRN, CVRN, SUPREME DM, etc.).
- Review descriptive information on the VDW data structures to identify available data for preliminary counts, as well as data elements that will be used for the research study itself. See additional VDW resources on our website.
- Review information about the HCSRN sites to consider potential relevant participants. See additional resources on our website.
- Find collaborating investigators at these sites. Discuss the data elements that are important for your study with the investigators and site programmers to understand their perceptions about the work that will be required to interpret each site's data.
- Obtain prep-to-research IRB approval.
- Work with your local site data manager to develop programs for extracting preliminary counts of critical data elements from the VDW at the participating sites. See the HCSRN Directory of Key Site Contacts.
- Some HCSRN sites have software tools that allow ad hoc cross-site queries with no additional programming effort. Examples: i2b2/SHRINE and HCSRNnet.
- Get boilerplate text, references and figures for the proposal. See Tools & Materials for proposal writing on our website.

## Using the VDW within a funded research project

- Complete the preliminary steps of IRB, subcontracting, and data use agreements (DUAs).
- Establish the data elements needed for the study via conference calls with site investigators and programmers. This is a critical step that will help to produce a smooth data collection process. You may also want to run preliminary tests to be certain data elements are clear. This is particularly important if you are using data from an early time period. As with any data source that was not developed specifically for research, there will be missing data and differences in the potential interpretations of some data elements across sites. Early and intense discussions and tests will elicit this information and allow you to construct a clean analytic dataset.

- Many HCSRN programmers and data managers are very knowledgeable about data elements in the VDW and common issues that arise. Take full advantage of their expertise!

- Select the central project programmer, typically a programmer at the lead site who is familiar with the VDW.

- Sometimes this is a programmer from a participating site with more experience using the VDW.

- Work with the central project programmer to develop the data extraction programs.

- As you design the process for extracting and merging data, consider what is minimally necessary to complete the study.

- The data extraction program sent to the sites can directly extract and transfer data to the central project site for analysis or the program can be constructed to calculate site-level results and transfer them to the central project site to be merged. The latter is safer since you do not transmit any individual level data, but it is less flexible and not always sufficient for the analytic demands of the project.

- Test the program at one of the participating sites and revise as necessary.

- Distribute the program across the sites.

- Transfer the extracted data to the central project programmer for testing and merging results or for producing analytic datasets.