VDW Boiler Plate – Short Version

The Cancer Research Network (CRN) maintains a data resource utility called the Virtual Data Warehouse (VDW). It was created as a mechanism to produce comparable data across sites for purposes of proposing and/or conducting research. The VDW is "virtual" in the sense that the raw ("real") data remain at the local sites; the VDW is not a multi-site physical database at a centralized data coordinating center. At the core of the VDW are a series of standardized file definitions. Content areas and data elements that are commonly required for research studies are identified, and data dictionaries are created for each of the content areas, specifying a common format for each of the elements—variable name, variable label, extended definition, code values, and value labels. Local site programmers have mapped the data elements from their HMO's legacy data systems onto this standardized set of variable definitions, names, and codes, as well as onto standardized SAS file formats. This common structure of the VDW files enables a SAS analyst at one site to write one program to extract and/or analyze data at all participating sites. The program from one site is emailed to programmers at other CRN sites to run against their own VDW files and the resultant de-identified data are transferred to the analytic site via a secure encrypted Web site.

As of Fall 2005, the standardized *content areas* that have been developed are:

- Enrollment
- Demographics
- Tumor
- Pharmacy
- Diagnoses and Procedures
- Census
- Vital Signs

Examples of standardized *methods* include defining people with continuous enrollment in the health plan, and the Deyo implementation of Charleson comorbidity index.

VDW Boiler Plate – Extended Version

The CRN Scientific Data and Resources Core (SDRC) maintains the HMORN Virtual Data Warehouse (VDW) as a set of distributed standardized files at each CRN site. Local site programmers have mapped the data elements from their HMO's legacy data systems onto a standardized set of variable definitions, names, and codes, as well as onto standardized SAS file formats. The various VDW files can be linked to produce a comprehensive utilization record for each study patient at each HMO. This common structure of the VDW files enables a SAS analyst at one site to write one program to extract and/or analyze data at all participating sites. The program from one site is emailed to programmers at other CRN sites to run against their own VDW files and the resultant de-identified data are transferred to the analytic site via a secure encrypted Web site.

The CRN leadership conceived the Virtual Data Warehouse (VDW) as a mechanism to produce comparable data across sites for purposes of proposing and/or conducting research. The VDW is "virtual" in the sense that the raw ("real") data remain at the local sites; the VDW is not a

multi-health plan physical database at a centralized data-coordinating center. At the core of the VDW are a series of standardized file definitions. Content areas and data elements that are commonly required for research studies are identified, and data dictionaries are created for each of the content areas, specifying a common format for each of the elements—variable name, variable label, extended definition, code values, and value labels. This allows HMO Site Data Managers to construct comparable data sets using potentially different sources and formats. Our vision is that using the VDW to manage the interfaces between project data needs and health plans legacy information systems, and between project analysts and health plan programmers will become a "best practice" for all CRN research projects.

Data standardization involves the following steps: 1) specifying common variable names, labels, coding, and definitions; 2) writing programs to extract and convert variables stored in HMO legacy information systems to the common standards; 3) testing standardized data for consistency and accuracy; 4) standardizing methods by writing macros that are used across projects; and 5) teaching researchers and their analysts how to use the VDW to guide construction of analysis files for approved research projects.

The CRN Project Leaders designate and prioritize content areas would be useful for cancer research, such as tumor registry, enrollment, and utilization. With input from site researchers, members of the SDRC with content experience or interest discuss which data elements are commonly required for research studies and are likely to be found across health plans. The number of elements included in each content area must be large enough to be useful for research but not so large that creating and using the VDW becomes unwieldy. For example, enrollment information might include employer, benefit, and family relationship information and consist of dozens of fields. VDW elements selected for this content area were the more commonly needed fields of patient identifier, enrollment by month and year, and type of payer (e.g., Medicare). The data dictionary consists of variable names, definitions, and formats, as well as information relevant to the availability, reliability, and validity of each variable.

The standardized content areas that have been developed as of 2005 include Enrollment, Demographics, Tumor, Pharmacy, Diagnoses and Procedures, Census, and Vital Signs. Examples of standardized methods include defining people with continuous enrollment in the health plan, and the Deyo implementation of Charleson comorbidity index.

Each site data manager writes the necessary code to extract information from health plan data systems and convert it into a file that matches the standardized files as closely as possible. If any variable is not available from health plan systems, that information is added to the documentation for the standardized files. If a match is not straightforward, that information is also documented to alert potential future users of validity and consistency problems when using or pooling data from multiple plans. Once the appropriate legacy systems have been identified for a given content area, the site data managers, using site-specific operational definitions, are responsible for writing and testing programs to extract the raw data and convert it into the format prescribed by the data dictionary. As the source systems are likely to be different at each site,

these programs may be very different from site to site. They all, however, should yield comparable data. Depending on the content area and the data manager's familiarity with the content area, the time required for data extraction may vary from site to site. Quality or reasonableness checks should be done at each step to ensure that the content of the VDW is what is intended. Each site maintains a central repository of the finalized programs, along with documentation of special issues.

Possible references:

"Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity Index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992; 45: 613-619".

For the VDW macro we added CPT codes and a couple of procedures for Peripheral vascular disorder.

Hornbrook MC, Hart G, Ellis JL, et al. Building a Virtual Cancer Research Organization. J Natl Cancer Inst Monogr 2005;35:12-25.