THE INVESTIGATOR'S GUIDE TO CRN DATA ANALYSIS

BACKGROUND

The Cancer Research Network (CRN) facilitates data management, analysis, and scientific infrastructure at several U.S. integrated health care systems ("sites"). While data are collected and maintained independently by each site, a system of data standards and automated processes known as the Virtual Data Warehouse (VDW) has been established to facilitate consistent data analysis across the network.

The purpose of this document is to serve as a guide for investigators who wish to use and analyze data from CRN sites.

ANALYSIS PLANNING

The investigator should consider the following issues when planning an analysis of CRN data.

SITE SELECTION

When choosing CRN sites to include in an analysis or proposed study, the investigator should allow for the fact that sites differ in the availability and extent of historical data. Furthermore, while the VDW simplifies many aspects of CRN data analysis, the frequency with which data are processed and added to the VDW varies across sites, data elements, and data sources. Finally, the availability of some data elements differs across sites. The VDW Implementations Overview on the HCSRN Portal (login required) summarizes the years of data available and the frequency with which data tables are updated at each site.

REPORTING DELAYS

The investigator may need to adjust the period of interest for an analysis to accommodate delayed reporting of some data elements, especially data collected via linking to external registries. Although most VDW tables are regularly updated by participating sites, certain types of data are not available until several months after the associated events occur. For example, cancer registry data are used for most CRN data analyses; however, new cancers are not immediately reported to registries and thus are not immediately available in the VDW. The same principle applies to death-related data, as death reports may not be immediately available from state and local vital records departments.

HEALTH PLAN ENROLLMENT

The investigator should consider how patient enrollment factors into an analysis. The VDW contains data about persons both currently and formerly enrolled in the participating site's health plans. If an analysis involves, for example, counting repeated procedures or prescription fills over time, then the investigator may want to limit the population of interest to persons who were continuously enrolled in the selected health plan(s) during the period of interest. Furthermore, at some CRN sites the population of persons receiving care at the site's facilities ("patients") may not overlap entirely with the population of persons insured by that site's health care plan(s) ("enrollees"). Different sets of data may be available on patients versus enrollees, so the investigator should specify which population is of interest for a given study.

CANCER DIAGNOSES

CANCER REGISTRIES

Cancer registries are the primary source of incident cancer diagnosis data in the VDW. Every CRN site receives data from a different internal, local, or regional registry, each of which varies in completeness of data elements and geographic coverage. In the United States there are three major registry systems: the American College of Surgeons Commission on Cancer (CoC), the Centers for Disease Control and Prevention's National Program of Cancer Registries (NPCR), and the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program. In addition, some health care systems (including several CRN sites) have their own in-house registries. Different cancer registries may use different data collection systems; for example, CoC member registries follow the Facility Oncology Registry Data Standards (FORDS) Manual, while SEER registries are required to use the SEER Program Coding and Staging Manual. Fortunately, most registries have also adopted the North American Association of Central Cancer Registries (NAACCR) consensus coding standards to facilitate common data capture and comparison across different registry systems.

Data from cancer registries form the basis of each CRN site's VDW Tumor table, which compiles data on incident cancers diagnosed at each CRN site. Each record contains some or all of the following data elements for a newly diagnosed tumor:

- Patient demographics (e.g., date of birth, gender, race)
- Date of diagnosis
- Location (ICD-O-3 codes)
- Morphology/histology (ICD-O codes; version varies by diagnosis year)
    - ICD-O-1: 1976–1989
    - ICD-O-2: 1990–2000
    - ICD-O-3: 2001–Present
- Stage (various staging systems available; see Cancer Staging section below)
- Anatomic site-specific factors (e.g., tumor markers)
- Dates of first treatments (e.g., chemotherapy, radiation, surgery; see Cancer Treatment section below)

The Tumor section of the VDW specifications provides a complete list of all data elements available in the VDW Tumor table. However, the investigator should keep in mind that not all data elements will be complete for every tumor and/or CRN site.

HEALTH PLANS

Some cancer diagnoses can also be identified in patient medical records and administrative claims data via ICD-9-CM diagnostic codes. However, the investigator should be aware that many diagnoses identified in this way represent prevalent cancers. Tumor registry data are the preferred source for incident cancer diagnoses.

CANCER STAGING

Many analyses of cancer-related data involve limiting and/or stratifying populations by cancer stage. The VDW Tumor table contains data (when obtained from cancer registries) on the following staging and data collection systems:

- The **SEER Summary Staging System** describes the spread of cancer relative to the primary organ affected. The system was initially established in 1977 to facilitate epidemiologic analysis of cancer control efforts over time. As such, it is still a required data element for reporting to most cancer registries; however, it is not often used by physicians, who typically use more detailed classification systems in clinical practice (National Cancer Institute, 2012).

- The **AJCC (American Joint Committee on Cancer) TNM Staging System** summarizes the size and/or extent of the primary tumor, degree of spread to nearby lymph nodes, and presence of metastasis (American Joint Committee on Cancer, 2010). It is the standard staging system used by physicians (American Joint Committee on Cancer, 2013). The system has evolved over its nearly 40-year history and is currently in its seventh edition (American Joint Committee on Cancer, 2010).

- The **SEER Extent of Disease (EOD) Coding Scheme** used 10-digit codes to summarize size and/or extent of the primary tumor as well as lymph node involvement. Unlike the AJCC TNM system, SEER EOD covered all cancer sites (National Cancer Institute, 2012). In 2004 the SEER EOD system was replaced by the Collaborative Staging system, described below (American College of Surgeons, 2007).

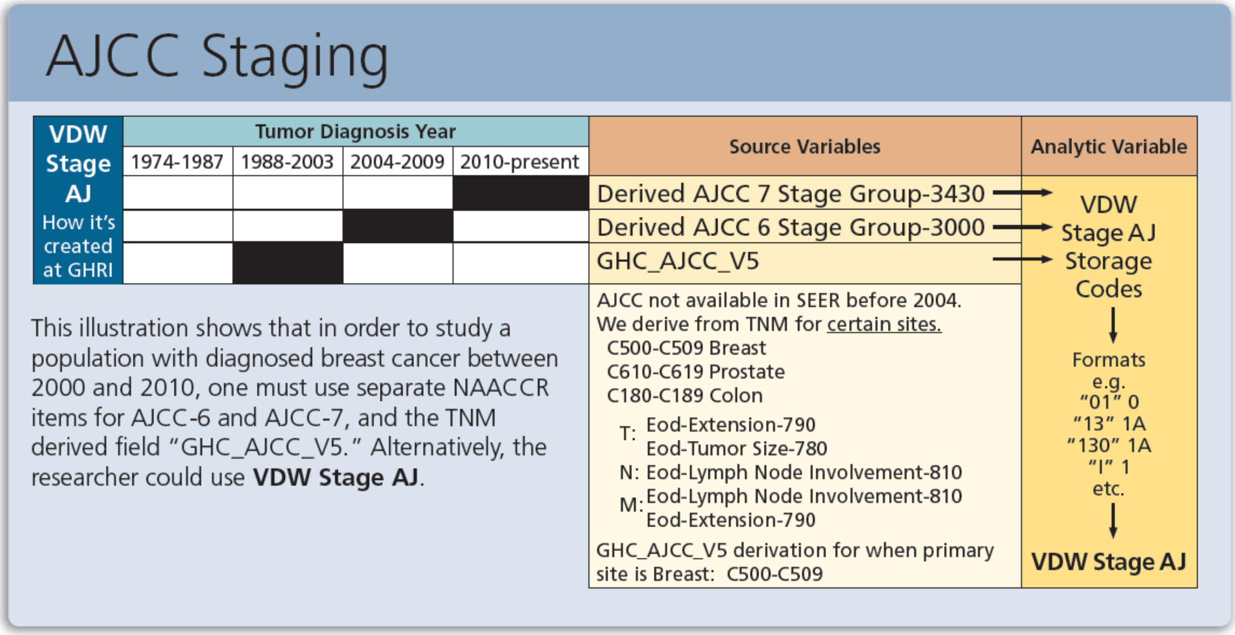
- The **Collaborative Stage Data Collection System (CS)** developed from the need to have a single system that would satisfy a variety of clinical and analytical purposes. Currently in its second version, the CS system uses a modified EOD format to encode various aspects of a given cancer. CS codes can then be translated into both AJCC TNM and SEER Summary stages (Collaborative Stage Data Collection System, 2012).

Table 1. Cancer Staging and Data Collection Systems Timeline

	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
SEER Summary	1977 Ed.																							2000 Ed.												
AJCC TNM		1st Ed.						2nd Ed.					3rd Ed.				4th Ed.					5th Ed.					6th Ed.							7th Ed.			
SEER EOD	13-Digit Codes						4-Digit Codes					10-Digit Codes																									
CS																												v1						v2			

As shown above, each of these staging systems has existed in several versions since its inception—a fact that complicates the analysis of cancer diagnoses across multiple years. Furthermore, cancer registry reporting requirements have changed over the years, meaning that different data elements were required by different registries at different points in time. As such, data completeness for VDW staging variables varies widely across years and sites.

NAACCR addressed the complexity of dealing with various cancer staging systems in the "Unresolved Issues" chapter of its Standards for Cancer Registries, Volume II: "The historic schemes were designed for different purposes at different times, and are not easily compared. . . . For these reasons, comparing cancer registry data by stage over time . . . is problematic" (Thornton, 2012). A poster on the CRN Portal (login required) prepared by Dustin Key and Jessica Chubak of Group Health Research Institute provides a glimpse at the complexity inherent in analyzing cancer staging data over time. The following excerpt specifically addresses AJCC TNM staging data at Group Health:



AJCC Staging

VDW Stage AJ How it's created at GHRI	Tumor Diagnosis Year				Source Variables	Analytic Variable
	1974-1987	1988-2003	2004-2009	2010-present		
				■	Derived AJCC 7 Stage Group-3430 →	VDW Stage AJ Storage Codes
			■		Derived AJCC 6 Stage Group-3000 →	
		■			GHC_AJCC_V5 →	

This illustration shows that in order to study a population with diagnosed breast cancer between 2000 and 2010, one must use separate NAACCR items for AJCC-6 and AJCC-7, and the TNM derived field "GHC_AJCC_V5." Alternatively, the researcher could use **VDW Stage AJ**.

AJCC not available in SEER before 2004.
We derive from TNM for certain sites.
C500-C509 Breast
C610-C619 Prostate
C180-C189 Colon
T: Eod-Extension-790
Eod-Tumor Size-780
N: Eod-Lymph Node Involvement-810
M: Eod-Lymph Node Involvement-810
Eod-Extension-790
GHC_AJCC_V5 derivation for when primary site is Breast: C500-C509

Figure 1. Excerpt from Key & Chubak Poster

In light of the issues described above, the investigator should plan to work closely with CRN programmers to determine the most appropriate staging variable(s) to use for a given analysis, taking into account clinical/scientific issues as well as data availability.

CANCER TREATMENT

The VDW contains data on cancer-related treatments, including surgery, chemotherapy, radiation, hormone therapy, and immunotherapy. These treatments can be identified in health plan data using a combination of ICD-9-CM diagnosis and procedure codes as well as prescription fill data. In addition, tumor registries collect some data on initial treatment of newly diagnosed cancers. Cancer treatment data are challenging to analyze due to frequent changes in and additions to relevant diagnosis, procedure, and prescription drug codes; as such, the investigator should plan to work closely with CRN programmers to determine the best ways of identifying cancer treatment data for a given analysis.

REQUESTING DATA

The investigator can request the use of CRN data for preparatory-to-research analysis for grant applications by downloading and completing the standard CRN Data Request Form. More detailed instructions on how to complete the form are linked here.

ADDITIONAL RESOURCES

CRN Website: http://www.hcsrn.org/crn/en

FORDS Manual: http://www.facs.org/cancer/coc/fordsmanual.html

NPCR Standards: http://www.cdc.gov/cancer/npcr/standards.htm

SEER Coding and Staging Manual: http://seer.cancer.gov/tools/codingmanuals/

NAACCR Standards: http://www.naaccr.org/

Cancer Staging Fact Sheet: http://www.cancer.gov/cancertopics/factsheet/detection/staging

SEER Summary Staging: http://training.seer.cancer.gov/staging/systems/summary/

AJCC TNM Staging: http://www.cancerstaging.org/

SEER EOD Coding: http://training.seer.cancer.gov/staging/systems/eod.html

Collaborative Staging: http://www.cancerstaging.org/cstage/index.html

WORKS CITED

American College of Surgeons. (2007, January 1). *Implementation Dates.* Retrieved from American College of Surgeons: Cancer Programs website: www.facs.org/cancer/coc/implementationdates.pdf

American Joint Committee on Cancer. (2010, May 5). *Publications & Electronic Products*. Retrieved from American Joint Committee on Cancer website: http://www.cancerstaging.org/products/pasteditions.html

American Joint Committee on Cancer. (2010, May 5). *What is Cancer Staging?* Retrieved July 30, 2013, from American Joint Committee on Cancer website: http://www.cancerstaging.org/mission/whatis.html

American Joint Committee on Cancer. (2013, April 15). *Staging Resources.* Retrieved from American Joint Committee on Cancer: http://www.cancerstaging.org/staging/needtoknow.pdf

Collaborative Stage Data Collection System. (2012, June). *CS Presentations.* Retrieved July 30, 2013, from Collaborative Stage Data Collection System website: http://www.cancerstaging.org/cstage/education/PDF/cs101-presentation.ppt

National Cancer Institute. (2012, November 27). *SEER Extent of Disease Coding*. Retrieved July 30, 2013, from SEER Training Modules website: http://training.seer.cancer.gov/staging/systems/eod.html

National Cancer Institute. (2012, November 27). *Summary Staging Defined*. Retrieved August 9, 2013, from SEER Training Modules website: http://training.seer.cancer.gov/ss2k/staging/

Thornton, M. L. (2012, June). *Standards for Cancer Registries Volume II, Chapter V.* Retrieved from North American Association of Central Cancer Registries website: http://www.naaccr.org/Applications/ContentReader/Default.aspx?c=5