Getting Started with the 2014 PUF

NCDB Participant Use File (NCDB PUF)

This document is designed to offer current and potential PUF investigators some basic guidelines and recommendations for how to approach the data provided in the NCDB PUFs. This document specifically addresses the PUF investigators as the primary reader.

Getting Started with the PUF Data Set

The Participant Use File (PUF) is pulled from the National Cancer Data Base (NCDB), a joint program of the American College of Surgeons Commission on Cancer (CoC) and the American Cancer Society (ACS), and is offered as an added value to clinical investigators at CoC-Accredited cancer programs who desire to conduct their own studies. The aim of the CoC and NCDB is to position investigators to successfully use the PUFs. There are a number of resources available to investigators. Chief among these is an on-line and publicly accessible PUF Data Dictionary (http://ncdbpuf.facs.org/), which has been developed as a resource to investigators using the NCDB PUF data files. Before proceeding with analyses, all investigators are encouraged to familiarize themselves with the on-line PUF Data Dictionary. This dictionary provides a wealth of information designed to facilitate the analytic use of the data items in the PUFs. This document is designed to expedite investigators' familiarization with the PUF data, and provides a list of the registry data items that should be considered when defining the analytic cohort of patients for proposed analyses. This introduction does not serve as a replacement for the on-line PUF Data Dictionary, which should be consulted before data analysis begins. Additionally, the NCDB provides a reference list of recent abstracts and publications that have been developed from PUF files, and categorized by cancer site.

Patients Included in the NCDB PUF Data Sets

Distributed PUFs are organ specific, based upon specific ICD-O primary site and histology combinations, and should be sufficient to address the study proposals described in reviewed applications. The site-histology combinations used to select the cases provided in the PUF data sets are documented: http://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html. Investigators are encouraged to run preliminary frequency distributions of all relevant data items, including primary site and tumor histology, in order to appropriately define the patient inclusion and exclusion criteria specific to their proposed project. The PUF data only include patient data from facilities that are currently accredited. VA and DoD facilities are excluded from PUF files. Facilities located in Puerto Rico are also excluded.

Please note that the NCDB PUF data are hospital based, NOT population based. Do not refer to the NCDB data as population based in any presentations or publications.

Reference to change in PUF file content due to active CoC Accreditation status:

If you had a previous PUF file for the same site as your new file, you may notice that the number of cases in the new file has decreased, especially for older diagnoses. The data you received this year are limited to cases reported by currently-accredited CoC hospitals. Cases reported by hospitals that are no longer accredited are excluded. Case reports for hospitals that are not currently accredited are not updated in the NCDB data, and their quality cannot be assured. The principal effect will be on cases more than 5-10 years old.

Data Items

Patient Identifiers

In compliance with the Health Insurance Portability and Accountability Act (HIPAA) regulations, PUFs have been stripped of all direct patient identifiers, de-identified according to the "Safe Harbor" rules. [1] The case identification number contained within is randomly assigned, and will change with each PUF version release. The PUF Case IDs are not the same across cancer sites, and cases cannot be linked across cancer sites. In the adult PUF file, pediatric patients have been excluded*, and patients 90 years of age or older are collapsed into the age group 90+ to maintain confidentiality. Facility Location and Facility Type are suppressed for cases aged 0-39. All dates have been removed and replaced with measures of elapsed time. The only date value appearing in the PUF is the year of diagnosis. The narrowest geographic unit available is that of the US Census Division in which the reporting hospital is located. Hospital identity has also been masked. For additional details, please see the Data De-Identification and Confidentiality document posted in the PUF on-line Data Dictionary. [2]

(*) Note: In the 2014 PUF data, there will be four PUF age cohorts: Pediatric (0-17), pediatric/young adult (0-39), adult (18-90+), and all ages (0-90+).

Year of Initial Diagnosis

The 2014 PUF Release includes data for patients diagnosed 2004 through 2014. The year of diagnosis should be used to select patients appropriate to the timeframe of the planned analysis. The availability of some data items is determined by diagnosis year, and not all data items in the PUF are available throughout the entire ten-year span of the PUF. To verify availability of data items by diagnosis year, be sure to review the description of each item in the on-line Data Dictionary.

NCDB PUFs		
File Name	**Diagnosis Years Included**	**Date of Release**
α Test Release	1998 – 2007	April 2010
βPUF Release	1998 – 2010	August 2012
2011 PUF Release	1998 – 2011	August 2013
2012 Semi-Annual PUF Releases	1998 – 2012	Fall 2014 and Spring 2015
2013 Semi-Annual PUF Releases	2004 – 2013	Fall 2015 and Spring 2016
2014 Semi-Annual PUF Releases	2004 – 2014	Fall 2016 and Spring 2017

PUF Multiple Source Item

All CoC accredited programs that initially diagnose a patient or that provide all or part of first course treatment report the case to the NCDB. If more than one facility submitted a report, the “best” is provided in the PUF file (PUF_MULT_SOURCE variable, coded 1), based on the most recent patient contact with the program, completeness of coded detail and/or edit quality, where differences exist. The record used in the case of ties is arbitrary. If this item is coded 0, only one facility provided a report for this cancer. This item should be used for hospital level comparisons using surgical volume, treatment, distance, or other hospital level computations in order to take into account cases treated at more than one hospital. Researchers can choose to limit hospital level analyses to only cases that received treatment at one CoC facility, or may choose to only include variables that indicate treatment was received at the facility included in the PUF. Researchers are encouraged to consult with NCDB staff to further clarify any questions regarding duplicate records and treatment in more than one facility.

Reference Date Flag

Every facility has a reference date, from which they are accountable for the completeness of the data for cases diagnosed in that year through the present. Since a facility may request to move their reference date forward, there are some instances where a case’s diagnosis year falls before the facility’s reference date. This item, REFERENCE_DATE_FLAG, is coded 0 in cases where this occurs. A 1 signifies cases where the diagnosis year is on or after the reference date year. Reports for cases whose diagnosis date is prior to the reference date cannot be changed or updated by the facility. For this reason, PUF researchers may choose to omit cases where the diagnosis date precedes the reference date, depending on the nature of the study. Note that, depending on diagnosis year and cancer site, excluding cases with diagnosis year preceding the reference year may omit greater than 40% of cases.

Sequence Number

The data item Sequence Number refers to the sequence of malignant and non-malignant tumors diagnosed in a patient and is used to distinguish cases with multiple cancer diagnoses. By default, your PUF includes all sequence codes available for each reported patient. Patients with only one lifetime cancer diagnosis will have a sequence number code value of 00. Sequence number 01 indicates that the reported tumor is the first of multiple diagnoses. The NCDB has no mechanism by which to link separate case reports of the same patient. It is customary to limit analyses to patients with sequence numbers 00 and 01 to ensure that any review of treatment or outcomes of the study cohort is not confounded by treatment administered for a prior cancer diagnosis. It is not uncommon to encounter high sequence numbers, especially among melanoma patients.

Behavior

The PUF includes in situ or non-invasive (behavior code 2) and malignant or invasive (behavior code 3) primaries. Non-malignant or borderline cases (behavior codes 0 and 1) are only available for primaries of the intracranial and central nervous system tumors.

Class of Case

The PUF only includes "analytic cases" whose initial diagnosis and/or treatment were/was performed at the reporting facility. Class of Case 00 denotes cases diagnosed at the reporting facility that did not receive any treatment at that facility. Class of Case 10-14 are cases that were initially diagnosed and provided all or part of their treatment at that facility. Class of Case codes 20-22 are those patients that were diagnosed at another facility and received all or part of their treatment at the reporting facility. If the focus of a proposed project is treatment, it would be standard practice to exclude Class of Case 00 cases from your study cohort. The Commission on Cancer does not require follow-up for Class of Case 00 cases, so they also should be excluded from survival studies.

Cancer Program Category

The PUF is limited to cancer programs currently accredited by the CoC. Facility Type provides a general classification of the reporting facility's structural characteristics, and defines a portion of the criteria required for CoC Accreditation. PUFs identify reporting facilities as one of four types: Community, Comprehensive, Academic/Research hospitals, or Integrated Network Cancer Programs. These categories follow the classification scheme used by the CoC accreditation program, and are determined by a variety of factors. If you are including facility type in your analyses, be aware that facilities in the Integrated Network Cancer Program (INCP) category are comprised of many types of facilities (such as Academic, Community, etc.), but are assigned the INCP type in the PUF data when they join a Network. . Cases reported from Veterans Affairs and Department of Defense facilities are not included in PUF data files. For additional details on data that are suppressed or cases that are omitted, please see the Data De-Identification and Confidentiality document posted in the PUF on-line Data Dictionary. This item is suppressed for cases aged 0-39.

Census and Urban/Rural Data

Area-based or environmental measures of patient income and education are provided in the PUF. These measures are derived by linking the reported ZIP code of the patient's residence at the time of diagnosis to year 2000 Census data. The data describing median household income and level of educational attainment represent the ZIP code of patient residence, not that of individual patients. Since the Census only uses the short form as of 2010, the majority of the information traditionally collected by the decennial census is now collected in the American Community Survey (ACS). The PUF will include the most recent ACS data released as of April 2014, which consists of survey years 2008-2012. The 5-year datasets are not just an average of each year in the period; the final estimate uses several weighting methods, among other adjustments. Items added are 2008-2012 median household income quartiles (MED_INC_QUAR_12) and 2008-2012 percent without high school degree quartiles (NO_HSD_QUAR_12.) The descriptions in the respective entries in the PUF data dictionary for each data item briefly describe the cautions of comparing with Census 2000, and more information can be found at: https://www.census.gov/acs/www/guidance_for_data_users/comparing_2012/. The data are extracted from the American Fact Finder website: http://factfinder.census.gov/.

The 2013 Rural-Urban Continuum data have also been added to the PUF. The 2003 Rural-Urban data are still included in the PUF, and the labels for the classification codes are the same in the 2003 and 2013 data, so a direct comparison may be made. More information can be found on the United States Department of Agriculture (USDA) website: http://www.ers.usda.gov/data-products/rural-urban-continuum-codes.

Comorbid Conditions

Comorbid disease burden is represented in the PUF as a summary value. This value is based on the Deyo adaptation (1992) of Charlson's comorbidity index and can be used as a mechanism to control for pre-existing medical conditions that may affect treatment decisions. The scores are mapped from as many as ten reported ICD-9-CM secondary diagnosis codes and are summed to create one value for each case, categorized as a total score of 0, 1, and 2 or more.

AJCC Stage of Disease

The AJCC clinical and pathologic stage groups included in the PUF are a TNM-based system coded or reported according to the edition corresponding to the patient's diagnosis year. The fifth edition of the AJCC Staging Manual is used to represent patients' cases diagnosed from 1998 through 2002. The sixth edition describes the anatomic extent of disease for patients diagnosed from 2003 through 2009. Patients diagnosed in 2010 are staged according to the seventh edition of the AJCC Staging Manual Data. **Exercise caution when using staging information.** Staging definitions may change between editions of the AJCC staging manuals, and rules delineating "stageable histologies" have become more specific over time, beginning with the 5th edition of the staging manual. Investigators may observe variability with the completeness of reported staging information over time. There have also been changes in CoC program standards and NCDB data reporting requirements in recent years. Rules defining which personnel within a facility were authorized to assign stage and the extent to which registry staff were directed to copy the recorded stage information have been both tightened and relaxed over the years captured in the PUF. In addition, the universal implementation of the Collaborative Stage Data Collection System (CS) in 2004 across all cancer registries in the United States contributed to a reduction in the coding of physician-reported staging in subsequent years.

Site Specific Factors from the Collaborative Stage Data Collection System

Several PUF projects examine one or more laboratory prognostic indicators. These are available as Site Specific Factors (SSF) collected as part of the Collaborative Stage Data Collection System (CS). The term "collaborative" means that the data collection tool was devised to meet the various needs of cancer registry data standard setters such as the Commission on Cancer (CoC), Surveillance Epidemiology and End Results (SEER), and the National Program of Cancer Registries (NPCR).

Up to 25 data fields are used to collect SSFs. Being site specific, they contain different information depending on the type of cancer in the report. For example, for breast cancer reports, SSF1 contains "Estrogen Receptor (ER) Assay" results, but for colon cancer reports, SSF1 contains "Carcinoembryonic Antigen (CEA)" results.

SSFs also may convey non-laboratory site-specific information that is relevant to prognosis for some cases. For example, SSF1 for gastric cancers is "Clinical Assessment of Regional Lymph Nodes", and for melanoma of skin it is "Measured Thickness (Depth), Breslow Measurement".

Some detective work is required to identify the data fields of interest, the applicable codes, and the adequacy of the data for the particular study:

I. The codes, and occasionally the fields used, for a particular prognostic factor have changed over time. In the PUF, the SSF data are retained in the form in which they were submitted. That means that it will be necessary to identify the CS Version Numbers that are used in the PUF file, and use those to identify whether the data contents for the desired SSF may have changed or moved. Links to the site-specific codes can be found at http://ncdbpuf.facs.org/?q=node/370. The CS web sites are maintained by Collaborative Stage Work Group of the American Joint Committee on Cancer. Select the applicable CS Version from the link above, and then select the schema name that applies to the cases in the project.

II. The quality of the SSF data items has undergone minimal review by NCDB, and PUF users are advised to examine the data consistency and completeness of these items carefully before proceeding with the study.

a. All SSF data items are edited for validity and internal consistency before the case report is submitted, and the submitter is required to correct any edit errors. However, some coding errors remain.

b. Case coverage of the SSFs is limited for a variety of reasons, potentially seriously affecting their applicability for some studies.

i. The availability of the measures to hospital registrars at the time of data entry is sparse for many prognostic measures. The source of information is usually the laboratory report as it appears in the hospital patient record. The information may not be available in the hospital if it was requested by a physician and the report was sent to the physician's office. Alternatively, it may be delayed and not picked up later.

ii. The individual tests are not run at all locations or for all patients, even if the test is part of an acknowledged treatment protocol.

iii. Finally, many hospital registries begin abstracting data for the years the measures were introduced prior to the hospital's upgrade of the software essential to collecting those items, and they did not necessarily return to the cases to abstract the missed data. Some of the SSFs were first introduced in 2004, and are underrepresented for cases diagnosed that year compared to later years. Most prognostic SSFs were introduced in 2010, and are certainly underrepresented for 2010 diagnoses; those are not available at all for earlier years.

The SSFs, the versions in which they were implemented, and whether the field was required for CoC registries are described in CoC SSFs for v0205.xlsx. As noted above, the fields in which these items were stored and the codes used may have changed over time. Spreadsheets for earlier versions of CS may be accessed via this link: http://seer.cancer.gov/csreqstatus/application.html (make sure to choose 'CoC' as the Standard Setter).

Treatment

The treatment information provided in the PUF is limited to "first course of treatment", which is defined as all methods of treatment recorded in the treatment plan and administered to the patient before disease progression or recurrence. "No therapy" is a treatment option that occurs if the patient refuses treatment, the family or guardian refuses treatment, the patient dies before treatment starts, or the physician recommends no treatment be given. Treatment plans describe the type(s) of therapies intended to modify, control, remove, or destroy proliferating cancer cells. Cancer registries are directed to review documentation confirming a treatment plan which may be found in several sources. Examples include medical or clinic records, consultation reports, and outpatient records. In addition:

- All therapies specified in the physician(s) treatment plan are a part of the first course of treatment if they are actually administered to the patient.
- A discharge plan must be part of the patient's record in a Joint Commission-approved program and may contain part, or all, of the treatment plan.
- An established protocol or accepted management guidelines for the disease can be considered a treatment plan in the absence of other written documentation.

- If there is no treatment plan, established protocol, or management guidelines, and consultation with a physician advisor is not possible, use the principle: "initial treatment must begin within four months of the date of initial diagnosis."
- The first course of treatment includes all therapy planned and administered by the physician(s) during the first diagnosis of cancer. Planned treatment may include multiple modes of therapy and may encompass intervals of a year or more. Any therapy administered after the discontinuation of first course treatment is subsequent treatment, and is not reported to the NCDB.
- The variable RX_SUMM_TREATMENT_STATUS indicates whether patients received any treatment or are under **active surveillance**. This variable was implemented in 2010 and is not available for prior diagnosis years.
- **Starting with the 2012 PUF (first released in 2014), additional treatment variables are now included, which indicate whether treatment was received at the reporting facility included in the PUF**. Versions prior to the 2012 PUF only included summary treatment variables, which indicated whether treatment was received at any CoC accredited facility, including facilities not included in the PUF data (see PUF Multiple Source Explanation above). The new treatment variables include “Surgery at this Facility”, “Chemotherapy at this Facility”, and “Other Treatment at this Facility”.

Distance Metrics

The PUF includes a "crow-fly" or great circle distance measure between the latitude and longitude of the centroid of patient's ZIP code of residence and the latitude and longitude of the facility mailing address. The precision of this item as an indicator of the true distance between two points is dependent upon the spatial area of the ZIP code and the proximity of the facility’s administrative mailing address to the actual treatment center**. If facility level analyses of distance to treatment are conducted using this variable, researchers need to take into account whether the treatment occurred at the facility included in the PUF data, or at a facility not included in the PUF.**

Outcomes

The CoC accreditation standards require an annual 90% follow-up rate for all living, eligible, analytic patients diagnosed within the last 5 years and an 80% follow-up for all eligible analytic cases from the cancer registry's reference date. Participating registries report patient follow-up to the NCDB annually. The PUF data do not include cause of death information, so cause-specific survival cannot be calculated. It is recommended that survival analyses be restricted to patient cohorts with only one reported cancer diagnosis (Sequence Number 00) in order to avoid confounding outcomes with patients who may have been diagnosed and treated for a separate malignancy. Vital Status information is not included for patients diagnosed in 2014 due to the limited follow up for these patients.

The PUF data also include both 30 and 90 day mortality for patients undergoing surgical resection (Surgery Primary Site Codes 20-90). If analyzing these data items at the facility level, the researcher needs to limit cases to those for whom the surgery was performed at the facility, using the variable Surgery at this Facility. The 30 and 90 day mortality items also do not include data for 2014 diagnoses, due to limited follow up.

If calculating survival, note that you may not publish or present any PUF data that compares your facility’s survival to the survival in the PUF. For more information on this policy, please refer to the letter sent to Commission on Cancer facilities in February 2015: **https://www.facs.org/~/media/files/quality%20programs/cancer/ncdb/coc_survivalpublic%20reporting%20policy.ashx.**

User Defined Data Items

Hospital Volume

Beginning in 1998, all CoC-accredited cancer programs are required to submit case reports to the NCDB in response to the annual Call for Data. In the NCDB PUF, facilities are assigned a random ID, *PUF_FACILITY_ID*. This ID is assigned regardless of cancer site, so researchers may identify the same facilities across cancer sites. The number of CoC-accredited cancer programs changes from one diagnosis year to the next. Thus, not all of the hospitals available in the PUF have been accredited for every one of the diagnosis years included in the PUF. If a planned analysis includes the calculation of hospital volume, investigators should recognize that CoC reporting requirements affect the methodological approaches to computing and estimating hospital volume.

Some patients receive treatment in more than one CoC accredited program, and are noted in the **PUF Multiple Source** item; however only one of the cancer programs where diagnosis/and or treatment was received is included in the PUF, in order to exclude duplicate records for the same patient. Thus calculation of hospital volume will not include hospitals with duplicate records that were excluded from the PUF. Approximately 10% of patients in the PUF data have duplicate records, but this percentage varies by cancer site. In addition, the summary surgical procedure of the primary site variable is based on information from the hospital where the surgery was performed, but this is not necessarily the hospital report that is included in the PUF data. The *at this facility* surgery variable may be used to account for surgery performed at that particular facility included in the PUF.

Whether an analysis uses total case volume or surgical volume, the easiest way to begin assigning average volume to each hospital appearing in the PUF is to create an aggregated dataset of the number of cases by hospital and diagnosis year. Such an aggregated file can be used to assess a particular cancer program's CoC accreditation history. Hospitals that have remained accredited throughout the years covered by the PUF pose minimal challenges when attributing volume metrics. If there are observed trends (either upward or downward) or spikes in hospital case counts, investigators may deem it more appropriate to calculate an average volume from the most recent years or a select set of years. Where significant shifts in annual caseloads are observed, investigators might consider recalculating their volume metric using a minimum and maximum volume value for each hospital in the aggregated dataset.

Hospitals that have previously discontinued and subsequently re-established their CoC accreditation throughout the span of diagnosis years available in the PUF will display a seemingly inconsistent or incomplete reporting pattern across years. Investigators should be certain to check their aggregated data set to ensure that computed volume metrics appropriately account for these hospitals.

In addition, researchers need to only include surgeries performed at the facility included in the PUF, by using the variable, Surgery at this Facility.

Citing Data from the NCDB

While citation is largely at the discretion of the author(s), there are three key components of information that must be conveyed on all peer-reviewed publications that draw from NCDB data:

- The NCDB is to be cited as a joint project of the American Cancer Society and the Commission on Cancer of the American College of Surgeons.
- The American College of Surgeons has executed a Business Associate Agreement that includes a data use agreement with each of its Commission on Cancer accredited hospitals.
- The NCDB, established in 1989, is a nationwide, facility-based, comprehensive clinical surveillance resource oncology data set that currently captures 70% of all newly diagnosed malignancies in the US annually.

A list of published studies using the NCDB is posted on-line at: http://www.facs.org/cancer/ncdb/biblclin.html. In addition, projects that have been completed using the NCDB PUF are posted on the PUF on-line Data Dictionary link at http://ncdbpuf.facs.org/?q=node/304.

These studies should serve to demonstrate how previous users have approached the use of NCDB data, the patient socio-demographic factors, staging information, or treatment data. These bibliographic and project lists may serve as a useful reference as PUF Principle Investigators begin to familiarize themselves with NCDB PUFs.

Ancillary Data References

- Deyo RA, Clerkin DC, Ciol MA. Adapting a clinical comorbidity index for use with the ICD-9-CM administrative databases. J Clin Epidemiol. 1992;45:613-619.
- AJCC 5th Edition Cancer Staging Manual. Fleming ID, Cooper JS, Henson DE, Hutter RVP, Kennedy BJ, Murphy GP, O'Sullivan B, Sobin LH, Yarbro JW (eds.). Lippincott-Raven Philadelphia, 1997.
- AJCC 6th Edition Cancer Staging Manual. Greene FL, Page DL, Fleming ID, Fritz A, Balch CM, Haller DG, Morrow M (eds.). Springer, New York, 2002.
- AJCC 7th Edition Cancer Staging Manual. Edge SE, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A (eds.). Springer, New York, 2009.
- Collaborative Staging Manual and Coding Instructions, Version 1.0. Collaborative Staging Task Force of the American Joint Committee on Cancer. NIH Publication Number 04-5496. US Department of Health and Human Services, National Institutes of Health, National Cancer Institute.
- Collaborative Stage Work Group of the American Joint Committee on Cancer. Collaborative Stage Data Collection System User Documentation and Coding Instructions, version 02.04.40. Published by American Joint Committee on Cancer (Chicago, IL).
- Facility Oncology Registry Data Standards (FORDS): Revised for 2013. Phillips JL, Stewart AK (eds.) American College of Surgeons 2002.
- International Classification of Diseases for Oncology (ICD-O), Third Edition. Percy C, Shanmugaratnam K, Whelan S, Parkin DM, Jack A, Fritz A, Sobin L (eds.) World Health Organization, Geneva, 2001.
- Registry Operations and Data Standards (ROADS), Standards of the Commission on Cancer Vol II. Johnson C (ed.). American College of Surgeons, 1998.

1. http://privacyruleandresearch.nih.gov/research_repositories.asp

2. http://ncdbpuf.facs.org/?q=node/275
3. http://www.facs.org/cancer/coc/whatis.html
4. http://www.facs.org/cancer/coc/programstandards2012.pdf
5. Bilimoria KY, Stewart AK, Winchester DP, Ko CY. The National Cancer Data Base: A Powerful Initiative to Improve Cancer Care in America. Ann Surg Oncol. 2008; on-line, Jan 9.