PROSPR II LOTUS/PROSPR-lung PUDS Data Dictionary

# Overall Instructions

In recognition of the need to improve lung and other types of cancer screening, in 2018, the National Cancer Institute (NCI) funded the large multisite consortium, Population-based Research to Optimize the Screening PRocess (PROSPR). The overall aim of PROSPR is to conduct multi-site, coordinated, transdisciplinary research to evaluate and improve cervical, colorectal, and lung cancer screening processes. Within the Lung PROPSR Research Consortium, titled Lung cancer screening Optimization in The United States (LOTUS), the aims are to assess utilization and outcomes of lung cancer screening across five diverse community-based healthcare systems and to identify key challenges and opportunities to optimize the impact of lung cancer screening. LOTUS has five data-contributing sites from diverse healthcare settings – 2 non-profit integrated health care delivery systems that provide insurance coverage and health care to members; 2 mixed-model healthcare systems providing coverage to enrolled members in facilities that are owned and operated by the healthcare system and providing health care to those with other types of insurance coverage; 1 academic health system providing primary care and other health care to patients with a variety of insurance products. LOTUS harmonized and collected data from the contributing sites to create a de-identified public use dataset (PUDS) for dissemination to the National Cancer Institute contractor managing dissemination to approved external collaborators. The PROSPR II LOTUS PUDS contains data on the LOTUS cohort who received primary care within, or who were enrolled members of, one of the integrated care delivery systems from January 1, 2017 through December 31, 2018, were ages 55-80 years as of January 1, 2017, and had complete data on smoking status on or before December 31st, 2017.

This data dictionary provides information and guidance for using the LOTUS/PROSPR-Lung PUDS in analyses. The LOTUS/PROSPR-Lung PUDS should be considered for pilot use only and is not appropriate for publication due to the inability to adjust for site in analyses. Data users wishing to analyze data for publication should apply to use a limited data set through the PROSPR DataShare site (<https://healthcaredelivery.cancer.gov/prospr/datashare.html>). Appropriate use-case scenarios for the LOTUS PUDS include the following:

1. Descriptive analyses for preliminary data needed for grant submissions;
2. Descriptive analyses to determine value in pursuing collaboration with PROSPR; and
3. Descriptive analyses for an educational project

The LOTUS PUDS2 data dictionary is divided into the following sections:

1. Section 1: Overview – A summary of the LOTUS PUDS cohort entry and exit criteria;
2. Section 2: Data file description and variable specifications; and
3. Section 3: Appendices

# Section 1: Overview – Description of the LOTUS PUDS Cohort Criteria

## LOTUS PUDS Cohort Entry and Exit Criteria

### **LOTUS PUDS Cohort Definition**

LOTUS cohort members entered the LOTUS PUDS cohort on January 1, 2017, if they met the following criteria:

1. Were ages 55-80 years old as of January 1st, 2017; and
2. Enrolled/engaged in the LOTUS study cohort; and
3. Had no personal history of lung cancer prior to January 1st, 2017; and
4. Had known smoking status that was recorded prior to December 31st, 2017; and
5. For health system membership-based LOTUS study cohort members, had continuous health system membership from January 1st, 2017 through December 31st, 2018 (allowing for a 90-day gap in enrollment); or
6. For primary-cared based individuals in the LOTUS study cohort (i.e., those that were not health plan enrollees in one of the LOTUS health systems), had a primary care visit during the year prior to January 1, 2017.

LOTUS cohort members exited the LOTUS PUDS cohort at the earliest occurrence of either of the following:

1. Reached the end of the LOTUS PUDS cohort follow-up period through December 31st, 2018; or
2. Death date prior to December 31st, 2018

### General Notes:

**Days Since Reference (DSR) Variables**

To mask dates of services and diagnoses, but to allow for calculating the days between events, each participant in the cohort has a random reference date assigned which in turn is used calculate “days since reference” (DSR) variables.  These variables allow one to compare temporal events within a participant’s record.

For example, a chest CT DSR of 1000 and a procedure DSR of 1100 indicates the procedure occurred 100 days (1,100 – 1,000) after the chest CT.

# Section 2: Data Files Description and variable specifications

## participant File

### **Overview**

This file contains one record of static covariates (i.e., covariates that do not change over time) for every LOTUS PUDS cohort participant. This table contains demographic variables (race, sex, ethnicity), as well as history of a lung cancer screening low-dose CT prior to January 1st, 2017.

### **Record Structure**

One record per LOTUS PUDS cohort member.

Data are derived from data-contributing site electronic health records systems.

### General Notes

For those with a Lung-RADS value of "4", the additional modifier, "A", "B", or "X", was included whenever it was available in the radiology report; some were just categorized as "4" in the radiology report and are categorized as "4" in the data. Early scans with missing Lung-RADS may have been read with Fleischner or other classification system and Lung-RADS results information is missing for those scans.

There are 490,237 records in the full Participant file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element Text** | **Type** | **Length** | **Formatting and/or Permissible Values** | **LOTUS Field Name** | **Instructions and Notes** | **Order** |
| Unique participant ID | char | 13 | Unique individual identifier with L prefix + number. This identifier can be used to link records across files for the same individual. | participant\_id |  | 1 |
| Sex | char | 2 | 0=Female | 1=Male |99=Unknown | sex |  | 2 |
| Hispanic or Latino origin | char | 2 | 0=No | 1=Yes | 99=Unknown | hispanic\_origin |  | 3 |
| Race American Indian or Alaska Native | char | 2 | 0=No | 1=Yes | 99=Unknown | race\_ak\_native\_american\_indian |  | 4 |
| Race Asian | char | 2 | 0=No | 1=Yes | 99=Unknown | race\_asian |  | 5 |
| Race Black or African-American | char | 2 | 0=No | 1=Yes | 99=Unknown | race\_black |  | 6 |
| Race Native Hawaiian or Other Pacific Islander | char | 2 | 0=No | 1=Yes | 99=Unknown | race\_pacific\_islander |  | 7 |
| Race White | char | 2 | 0=No | 1=Yes | 99=Unknown | race\_white |  | 8 |
| Race Another | char | 2 | 0=No | 1=Yes | 99=Unknown | race\_another |  | 9 |
| Age at cohort entry date | num | 8 | Range: 55-80 | Unknown: blank | age\_at\_cohort\_entry | Age on 1/1/2017 | 10 |
| Prior LCS LDCT (G0297, S8032, or equivalent) | Char | 1 | 0= No | 1=Yes | has\_prior\_lcs\_ldct | Had at least one lung cancer screening low-dose CT prior to 1/1/2017 | 11 |
| Year of Last LDCT prior to January 1, 2017 | char | 5 | Date: YYYYY | Unknown: blank | year\_of\_prior\_lcs\_ldct | Year of last lung cancer screening low-dose CT prior to 1/1/2017 | 12 |
| Lung-RADS of Last LCS LDCT prior to January 1, 2017 | char | 5 | 0 = incomplete | 1 = negative | 2 = benign appearance or behavior | 3 = probably benign | 4 = suspicious | 4A = suspicious | 4B/4X = suspicious | 97 = Other | 99 = Missing | results\_of\_prior\_lcs\_ldct | Lung-RADS interpretation of the last lung cancer screening low-dose CT prior to 1/1/2017 | 13 |
| Vital Status | char | 1 | A: No Death Recorded D: Death Recorded | vital\_status\_at\_cohort\_cutoff | Vital Status at end of 12/31/2018. Only including Excellent Confidence deaths. | 14 |
| Death DSR | num | 8 | number | dsr\_at\_death | Days since reference at date of death. | 15 |
| Data dictionary version number to which the data corresponds | char | 3 | 1.0 | data\_dictionary\_version |  | 16 |

## COHORT ENTRY yEAR File

### **Overview**

This file contains one record for each participant in the LOTUS PUDS cohort and includes variables that may change overtime.

### **Record Structure**

One record per LOTUS PUDS cohort participant; each record pertains to the year of cohort entry (i.e., 2017) but may include data collected prior to 2017 (see General Notes).

Data are derived from data-contributing site electronic health records systems.

### General Notes

The value for each variable is the value based on electronic health records systems data that was last record on, or prior to, December 31st, 2017. If a temporal measure is not available in 2017, the study looks back to the nearest record. For example, if a value was last recorded prior to 2017, the value is recorded in file with the year that the value was last recorded noted in the file. Allowing values recorded prior to 2017 limits missing data for individuals that do not have a value recorded during the 2017 calendar year.

The Charlson comorbidity index uses diagnosis codes from in-person health system encounters occurring during the 365-day period spanning from January 1st, 2017 through December 31st, 2017. The following references provide additional information on Charlson comorbidity index calculation:

-Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45:613–19.

-Quan H, Sundararajan V, et al. Coding algorithms for defining Comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43(11):1130-9

There are 490,237 records in the Cohort Entry Year file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element Text** | **Type** | **Length** | **Formatting and/or Permissible Values** | **LOTUS Field Name** | **Instructions and Notes** | **Order** |
| Unique participant ID | char | 13 | Unique individual identifier with L prefix + number. This identifier can be used to link records across files for the same individual. | participant\_id |  | 1 |
| BMI | num | 8 | Numeric | Unknown = . | bmi |  | 2 |
| Year BMI Recorded | char | 4 | YYYY | year\_of\_bmi\_measure\_date |  | 3 |
| Age at BMI Date | num | 8 | Numeric | Unknown = blank | age\_at\_bmi\_measure\_date |  | 4 |
| Was the participant insured? | char | 2 | 1=Yes | 99=Unknown | is\_insured | At any time during the calendar year. | 5 |
| Charlson comorbidity index | num | 8 | Numeric | Unknown = . | charlson\_index |  | 6 |
| Smoking - Current Status | char | 2 | 00=Never Smoked | 01=Former | 02= Current | current\_smoking\_status | Smoking status | 7 |
| Year of Smoking Status | char | 4 | YYYY | year\_of\_smoking\_status | Year smoking status recorded | 8 |
| Pack Years | num | 8 | Range: 0 - 100 | Unknown= . | pack\_years | Pack years ceiling is 100. Values above 100 are reset to 100. If an individual is recorded as having never smoked, pack-years is set to 0. | 9 |
| Pack Years Recorded Year | char | 4 | YYYY | year\_of\_pack\_years\_measure\_date | Year that pack years was recorded | 10 |
| Age at Pack Years Recorded Date | num | 8 | Integer| Unknown = blank | age\_at\_pack\_years\_measure | Age of patient on the date that pack years was recorded | 11 |
| Quit Years | num | 8 | Integer | Unknown = blank | years\_since\_quit | Number of years since patient quit smoking as of the date that quit information was recorded | 12 |
| Age at Quit Measure Date | num | 8 | Integer | Unknown = blank | age\_at\_quit\_years\_measure | Age of patient on the date that quit information was recorded | 13 |
| Year of Quit Measure Date | char | 4 | YYYY | year\_of\_quit\_years\_measure\_Date | Year when the quit years information was recorded | 14 |
| Data dictionary version number | char | 3 | 1.0 | data\_dictionary\_version | Enter the version number of this Data Request Packet. | 15 |

## cancer registry file

### **Overview**

This file contains one record for each individual in the LOTUS PUDS cohort who was diagnosed with an invasive lung cancer from January 1st, 2017 through December 31st, 2018. In situ lung cancers are not included.

### **Record Structure**

One record per invasive lung cancer diagnosed January 1st,2017 through December 31st, 2018 within the LOTUS PUDS study cohort.

Data are derived from abstracted tumor registry data at each health system.

### General Harmonization Notes

Variables that are based on the North American Association of Central Cancer Registries (NAACCR) are noted with refence to the specific NAACCR Item #, below. Additional information of NAACCR data standards can be found at: [www.naaccr.org](http://www.naaccr.org).

There are 1,553 records in the full Cancer Registry file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element Text** | **Type** | **Length** | **Formatting and/or Permissible Values** | **LOTUS Field Name** | **Instructions and Notes** | **Order** |
| Unique participant ID | char | 13 | Unique individual identifier with L prefix + number. This identifier can be used to link records across files for the same individual. | participant\_id |  | 1 |
| Sequence number of primary tumor at this reporting facility | char | 2 | 00=One primary in the patient’s lifetime |  01=First of two or more primaries |  02=Second of two or more primaries |  03=Third of three or more primaries | 04=Fourth or greater primary | 88=Unspecified or unknown sequence number of non-malignant tumor or central-registry defined neoplasms |  99=Unspecified or unknown sequence number of Federally required in situ or malignant tumors | sequence\_number | NAACCR Item #380 | 2 |
| ICD Code, primary site | char | 4 | C340 | C341 | C342 | C343 | C348 | C349 | primary\_site | NAACCR Item #400. Records the primary site using the topography codes listed in the International Classification of Diseases for Oncology, Third Edition (ICD-O-3). | 3 |
| Cancer Diagnosis DSR | num | 8 | integer | dsr\_from\_cancer\_dx | Days since reference at invasive lung cancer diagnosis. | 4 |
| Tumor malignancy or behavior | char | 2 | 0=Benign | 1=Uncertain/borderline | 2=In situ | 3=Malignant/invasive | 99=Unknown | tumor\_behavior | NAACCR Item #=523 Records the malignant potential of the tumor, ranging from /0 benign to /3 malignant (invasive). For UPHS, malignant status only is captured (In-situ or Malignant). | 5 |
| Cancer stage (Best AJCC) | char | 3 | 0 | 1 | 1A | 1A1 | 1A2 | 1A3 | 1B | 2 | 2A | 2B | 3 | 3A | 3B | 3C | 4 | 4A | 4B | OC=Occult | 88=Not applicable | 99=Unknown | ajcc\_stage | Staging according to the American Joint Commission on Cancer Staging Manual. Reference: https://cancerstaging.org/cstage/coding/Documents/csmanual-p1s1.pdf, page I-100 | 6 |
| SEER Stage | char | 1 | 1: Localized | 2: Regional by direct extension | 3: Regional to lymph nodes | 4: Regional both direct extension and lymph nodes | 5: Regional, NOS | 7: Distant metastasis | 8: staging scheme not applicable | 9: Unstageable, unknown, unspecified | seer\_stage |  | 7 |
| Tumor histology or cell type | char | 1 | 1 = 'Squamous' | 2 = 'Small cell' | 3 = 'NSCLC' | 4 = 'Other' | histology |  | 8 |
| Data dictionary version number to which the data corresponds | char | 3 | 1.0 | data\_dictionary\_version | Enter the version number of this Data Request Packet. | 9 |

## YOST Index social determinant of health file

### **Overview**

This table contains the Yost Index quintile based on US National data and separately, on state-specific data, for each LOTUS PUDS cohort participant. The Yost Index national and state quintiles are based on the participant’s census tract of residence at cohort entry or on the first date after cohort entry where a valid address is recorded in the electronic health record.

### **Record Structure**

One record per LOTUS cohort participant.

Data are derived from electronic health record data on address history that is linked to census tract for each individual using GIS software.

### General Notes

The Yost Index is a composite score of seven variables that measure different aspects of socioeconomic status within a census tract. The variables include median household income, median house value, median rent, percent of tract population below 150% of the poverty line, Education index, percent working class, and percent unemployed.

Additional information on the Yost Index can be found at:

-Yu M, Tatalovich Z, Gibson JT, Cronin KA. Using a composite index of socioeconomic status to investigate health disparities while protecting the confidentiality of cancer registry data. *Cancer Causes Control*. Jan 2014;25(1):81-92. doi:10.1007/s10552-013-0310-1

-Boscoe FP, Liu B, Lee F. A comparison of two neighborhood-level socioeconomic indexes in the United States. *Spat Spatiotemporal Epidemiol*. Jun 2021;37:100412. doi:10.1016/j.sste.2021.100412

There are 490,237 records in the full Yost Index file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element Text** | **Type** | **Length** | **Formatting and/or Permissible Values** | **LOTUS Field Name** | **Instructions and Notes** | **Order** |
| Unique participant ID | char | 13 | Unique individual identifier with L prefix + number. This identifier can be used to link records across files for the same individual. | participant\_id |  | 1 |
| Year of Address Date | char | 4 | YYYY | year\_of\_address\_date |  | 2 |
| Yost Quintile (Total US) | num | 8 | Range: 1 -5 | Unknown = 99 | yost\_overall\_quintile |  | 3 |
| Yost Quintile (by state) | num | 8 | Range: 1 -5 | Unknown =99 | yost\_state\_quintile |  | 4 |
| Data dictionary version number | char | 3 | 1.0 | data\_dictionary\_version |  | 5 |

## procedure file

### **Overview**

This file contains one record for each procedure occuring January 1st, 2017 through December 31st, 2018. The list of possible procedures included in the LOTUS PUDS data set is in the Appendix.

### **Record Structure**

One record per LOTUS PUDS participant per procedure.

Data are derived from data-contributing site electronic health records systems.

### General Notes

Procedures are restricted to procedures occurring 2017 through 2018 evaluated in the following publication. All procedures with the procedure codes listed in Appendix are included, regardless of a participant’s prior lung cancer screening history and regardless of indication.

- Zhao H, Xu Y, Huo J, Burks AC, Ost DE, Shih YCT. Updated analysis of complication rates associated with invasive diagnostic procedures after lung cancer screening. *JAMA Netw Open*. 2020;3(12).e2029874.

There are 20,730 records in the full Procedure file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element Text** | **Type** | **Length** | **Formatting and/or Permissible Values** | **LOTUS Field Name** | **Instructions and Notes** | **Order** |
| Unique participant ID | char | 13 | Unique individual identifier with L prefix + number. This identifier can be used to link records across files for the same individual. | participant\_id |  | 1 |
| Procedure Code Type | char | 1 | '1' = 'Bronchoscopy procedures'  '2' = 'Cytology / needle biopsy procedures'  '3' = 'Thoracic lung surgical procedures'  '4' = 'Other surgical procedures' | procedure\_code\_type |  | 2 |
| Procedure DSR | num | 8 | integer | dsr\_at\_procedure | Days since reference at procedure date | 3 |
| Data dictionary version number | char | 3 | 1.0 | data\_dictionary\_version |  | 4 |

## CHEST CT FILE

### Overview

This file is restricted to LOTUS PUDS cohort participants who had at least one lung cancer screening low-dose CT from January 1st, 2017 through December 31st, 2018. It contains participants’ first lung cancer screening low-dose CT that occurred during the timeframe from January 1st, 2017 through December 31st, 2018 and every chest CT (regardless of indication) that occurs after that lung cancer screening low-dose CT through December 31st, 2018.

### Record Structure

One record per LOTUS PUDS participant per chest CT.

Data are derived from data-contributing site electronic health records systems.

### General Notes

This table only includes EHR derived chest CT and not claims based procedures.

After the initial lung cancer screening related CT occurring January 1st, 2017 through December 31st, 2018 for a LOTUS PUDS cohort participant, all chest CTs (codes: G0297 | S8032 | 71250 | 71260 | 71275) for that individual are included in this file, regardless of indication for chest CT.

For those with a Lung-RADS value of "4", the additional modifier, "A", "B", or "X", was included whenever it was available in the radiology report; some were just categorized as "4" in the radiology report and are categorized as "4" in the data. Lung-RADS may be missing due to the Lung-RADS classification system not being used for a particular chest CT or due to missing data in the electronic health record.

There are 7,022 records in the full Chest CT file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element Text** | **Type** | **Length** | **Formatting and/or Permissible Values** | **LOTUS Field Name** | **Instructions and Notes** | **Order** |
| Unique participant ID | char | 13 | Unique individual identifier with L prefix + number. This identifier can be used to link records across files for the same individual. | participant\_id |  | 1 |
| Chest CT Days Since Reference | num | 8 | integer | dsr\_at\_chest\_ct | Days since reference at chest CT date. | 2 |
| Chest CT Procedure Code | char | 10 | G0297 | S8032 | 71250 | 71260 | 71275 | chest\_ct\_procedure\_code |  | 3 |
| Result of Chest CT scan (Lung-RADS) | char | 5 | 0 = incomplete | 1 = negative | 2 = benign appearance or behavior | 3 = probably benign | 4 = suspicious | 4A = suspicious | 4B /4X = suspicious | 97 = Other | 99 = Missing | chest\_ct\_scan\_result |  | 4 |
| Indication for Chest CT scan | char | 2 | 1 = Screening | 99 = Unknown | indication | Screening = Procedure code (G0297, S8032) indicates this is a baseline or annual Low Density Chest CT (LDCT) scan  Other = All other Chest CT procedure codes. | 5 |
| Insurance coverage at time of Chest CT | char | 2 | 1=Yes | 99=Unknown | is\_insured |  | 6 |
| Data dictionary version number to which the data corresponds | char | 3 | 1.0 | data\_dictionary\_version | Enter the version number of this Data Request Packet. | 7 |

**appendix: list of possible procedure codes included in the Lotus PUDS procedure file, GROUP INTO MAJOR CATEGORIES ACCORDING TO ZHAO, ET AL. 2020**

**Source:**

Zhao H, Xu Y, Huo J, Burks AC, Ost DE, Shih YCT. Updated analysis of complication rates associated with invasive diagnostic procedures after lung cancer screening. *JAMA Netw Open*. 2020;3(12).e2029874.





















