NLST USER GUIDE

TABLE OF CONTENTS

1. DATASET INFORMATION GUIDE

INTRODUCTION

The sections below describe important basic information about the NLST datasets. Additional information is available from other resources on the CDAS web site, including the data dictionaries and the dataset descriptions on the Datasets web page.

STUDY YEAR TERMINOLOGY (T[X])

Time on study in NLST is often described in "study years" using the notation T[X], as in T0, T1, T2, etc. This represents the number of years completed since randomization. During the screening phase of the trial (T0 - T2), the study year changed on the date of each screening exam. For instance, T0 begins on the date of randomization and ends on the date of the T1 screening exam, which was approximately one year later. After the screening phase (T3 - T7), the study year changed on the date of the anniversary of randomization. The T[X] screening exams were expected to occur at the beginning of the T[X] study year.

IMPORTANT VARIABLES

The following variables from the participant dataset are essential for many analyses.

- PID: participant identifier; a numeric code that uniquely identifies each participant.
- RNDGROUP: study arm; indicates which screening exam (CT or X-ray) participants were assigned to receive.
- STUDY: trial component (LSS or ACRIN). Each participant's involvement in NLST was facilitated by one of 33 institutions (screening centers) that enrolled, screened, and recorded data about each participant. STUDY indicates whether the participant's screening center belonged to the LSS network of 10 centers or the ACRIN network of 23 centers.
- SCR_RES0-2: results from the T0 – T2 screening exams; indicates whether lung cancer was suspected at the participant's exam.
- CONFLC: confirmed lung cancer status; indicates lung cancer diagnosis at any time during the trial.
- CANDX_DAYS: days from randomization to diagnosis of lung cancer.
- FINALDEATHLC: indicates whether lung cancer was the official cause of death.
- FUP_DAYS: days from randomization to death (for participants who died during NLST) or last contact (for all other participants).

HOW TO COMBINE DATASETS

The participant dataset contains all information necessary for most standard analyses. However, more detailed information is available for certain trial events. Other datasets cover screening exams, abnormalities, diagnostic procedures, medical complications, additional lung cancer characteristics and

multiple primary tumors, treatments, causes of death, non-cancer conditions, and contamination assessment surveys. All datasets contain the personal identifier PID, which should be used when merging datasets.

The study year of events should also be used to combine datasets. The study year variables in different datasets represent different events. In some datasets, the variables represent the study year of screening. In others, they represent the study year of lung cancer diagnosis. In a few datasets, the study year variable corresponds to three different situations for different participants: (1) study year of screen only, (2) study year of cancer diagnosis only, or (3) study year of both screening and diagnosis. Case (3) only occurs when a participant's positive screen is followed by a cancer diagnosis in the same study year. The details for each dataset are described below. Also, the name of the study year variable is not the same in all datasets, so the variable names are included.

- Screening and Abnormality datasets: STUDY_YR is the study year of screen.
- Lung cancer dataset: STUDY_YR is the study year of cancer diagnosis.
- Treatment dataset: TREAT_YEAR is the study year of cancer diagnosis.
- Participant dataset
    - Study year of screen is implicit in variables with an index of 0 to 2 (such as SCR_RES0, 1, and 2).
    - CANCYR is the study year of cancer diagnosis.
- Diagnostic procedure dataset: PROC_YEAR is the study year of screen and/or diagnosis.
- Medical complication dataset: COMP_YEAR is the study year of screen and/or diagnosis.
- Non-cancer condition datasets: STUDY_YR is the study year of screen and/or diagnosis.

To combine the datasets for abnormalities and comparison read abnormalities for either the CT or X-ray arm, link the datasets using the variables PID, STUDY_YR, and either SCT_AB_NUM or XRY_AB_NUM, depending on which arm's datasets are being used.

ACRONYM GLOSSARY

Below is a list of common acronyms found in the NLST data and on the CDAS web site.

Acronym	Definition	Category	Use
ACRIN	American College of Radiology Imaging Network	Study	Group of 23 (out of 33 total) NLST screening centers.
AJCC	American Joint Committee on Cancer	Resource	Provides standards for staging. NLST used the 6th edition of the AJCC staging manual when collecting data; 7th edition data will be available in the future.
Bx	Biopsy	Medical	
CDQ	Cause of Death Questionnaire	Form	Captures data from the death review process (EVP).
CDAS	Cancer Data Access System	Web site	Web site that handles requests for NLST data. https://cdas.cancer.gov
CXR	Chest X-Ray	Screen	One of the two screening modalities used in NLST. Also abbreviated as XRY.
DICOM	Digital Imaging and Communications in Medicine	Resource	Standard for medical imaging data. Images from NLST's CT screening exams were stored in DICOM format and are available from TCIA.
Dx	Diagnosis	Medical	
EVP	Endpoint Verification Process	Study	Review of medical records of selected deceased participants to determine whether lung cancer was the cause of death.
HAQ	Health Assessment Questionnaire	Form	Annual survey at LSS centers to assess use of spiral CT, chest X-ray, and other medical procedures outside of the trial protocol.
ICD-10	International Classification of Diseases, 10th Revision	Resource	Standard disease classification system; used for all deaths that occurred during NLST.
ICD-O-3	International Classification of Diseases for Oncology, 3rd Ed.	Resource	Standard cancer classification system; used for all cancers diagnosed during NLST.
LSS	Lung Screening Study	Study	Group of 10 (out of 33 total) NLST screening centers.
MRA	Medical Record Abstraction	Study	Process used to obtain cancer data in NLST.
NDI	National Death Index	Resource	Comprehensive database of US deaths; queried yearly to ensure complete ascertainment of deaths.
NLST	National Lung Screening Trial	Study	
NOS	Not Otherwise Specified	Abbreviation	
Rx	Treatment	Medical	
SCT	Spiral CT	Screen	One of the two screening modalities used in NLST.
T[X]	Time Point [X]	Reference	Denotes years since randomization.
TCIA	The Cancer Imaging Archive	Web site	Web-based repository for imaging and related data from cancer research, including the NLST CT screening exams. https://www.cancerimagingarchive.net/
TNM	Tumor / Node / Metastasis	Medical	Three main features used to determine the cancer stage under the AJCC system.
XRY	Chest X-Ray	Screen	One of the two screening modalities used in NLST. Also abbreviated as CXR.

Table 1: Table of acronyms and their meanings

2. GUIDE FOR ANALYSES OF MORTALITY, SURVIVAL, AND INCIDENCE IN NLST

INTRODUCTION

In order to keep confidential the identity of participants in NLST, the datasets released to the research community do not contain data on the calendar dates of events; such data could be used to determine a participant's identity. However, the datasets contain variables for the number of days from study entry (at randomization) to each event, which permits time-based analyses to be performed.

Analyses of mortality, survival, and incidence on NLST data should have a calendar date cutoff of 12/31/09 for follow-up time, because the database includes all deaths and cancer diagnoses through that date but none beyond it. However, the follow-up time variable in the dataset (FUP_DAYS) extends to the date of last contact, which for most participants was a study update form collected early in 2010 at various dates. While analyses would be simpler if the follow-up time variable stopped exactly at 12/31/09, that would allow back calculation of dates for a majority of participants and defeat the purpose of removing dates from the datasets. Adjustments can be made to the follow-up time variable to produce results equivalent to an analysis cut off at 12/31/09; details are provided below.

An analysis with a cutoff of 1/15/09 for mortality is also possible with the use of the DEATHCUTOFF variable. This matches the time period for the lung cancer mortality analysis in the NLST primary results paper. It is not possible to do a proper analysis of incidence or survival through 1/15/09, because the datasets do not contain a variable to indicate whether a diagnosis of lung cancer occurred before 1/15/09.

Note that the datasets represent a newer database (with more complete data collection) than the one used for the primary results paper, so the numbers of events and the follow-up time are slightly higher in the dataset than in the published results. A comparison will be presented after the person-year adjustments are explained.

PERSON-YEAR ADJUSTMENT FOR 12/31/09 CUTOFF (MORTALITY / SURVIVAL)

An adjusted follow-up time variable can be used to produce analyses of mortality or survival with results very similar to an analysis conducted with full knowledge of dates. The adjusted variable is created by subtracting a fixed number of days from all alive participants, but not changing the follow-up time for deceased participants. For an analysis through 12/31/09, the numbers of days to subtract are 58.1772 for the CT arm and 58.7590 for the X-ray arm; the person-years of follow-up time for such an analysis agree with the actual person-years through 12/31/09. See the code sample and results below.

```
if deathstat=0 then do;  **alive;
     if rndgroup=1 then fup_days_adj = (fup_days – 58.1772); **CT arm;
     else fup_days_adj = (fup_days – 58.7590); **X-ray arm;
end;
else fup_days_adj = fup_days;  **deceased: no change;
```

	Actual PY through 12/31/09	Adjusted PY
Spiral CT	167462	167462
Chest X-ray	166384	166384
Total	333846	333846

Table 2: Actual and adjusted person-years of follow-up for a mortality analysis with a cutoff of 12/31/09

Note that FUP_DAYS stops before 12/31/09 for all deceased participants and for participants whose last contact occurred before that date. While it would be more sensible not to adjust the person-time for alive participants whose last contact was before 12/31/09, it is not possible to identify them systematically using the dataset.

The adjusted follow-up time is negative for alive individuals whose last contact was very early in the study; their time can be set to 0 without introducing much error (~ 17 person-years).

PERSON-YEAR ADJUSTMENT FOR 1/15/09 CUTOFF (MORTALITY)

A similar adjustment can be made to do an analysis of mortality through 1/15/09. This is the cutoff date used for the official final analysis of lung cancer mortality published in the NLST primary results paper. The adjustment factors in this case are 394.6020 days for the CT arm and 392.1746 days for the X-ray arm. To restrict death events to only those which occurred before 1/15/09, use the condition DEATHCUTOFF=1. See the code sample and results below.

```
if (DEATHCUTOFF=0 or DEATHCUTOFF=2) then do;  **alive;
      if rndgroup=1 then fup_days_adj = (fup_days - 394.6020); **CT arm;
      else fup_days_adj = (fup_days - 392.1746); **X-ray arm;
end;
else fup_days_adj = fup_days;  **deceased: no change;
```

	Actual PY through 1/15/09	Adjusted PY
Spiral CT	144149	144148
Chest X-ray	143394	143394
Total	287543	287542

Table 3: Actual and adjusted person-years of follow-up for a mortality analysis with a cutoff of 1/15/09

The adjusted follow-up time is negative for alive individuals whose last contact was very early in the study; their time can be set to 0 without introducing much error (~ 180 person-years).

PERSON-YEAR ADJUSTMENT FOR 12/31/09 CUTOFF (INCIDENCE)

For an analysis of lung cancer incidence through 12/31/09 (as in the NLST primary results paper), a similar adjustment can be made. For lung cancer cases (CONFLC=1), use the time to diagnosis (CANDX_DAYS). For everyone else, do the same adjustment as for the mortality analysis through 12/31/09, with the same factors: 58.1772 for the CT arm and 58.7590 for the X-ray arm. See the code sample and results below.

```
if conflc=1 then lc_exit_days= candx_days; **lung cancer case;
else if deathstat=0 then do;  **alive;
     if rndgroup=1 then lc_exit_days = (fup_days - 58.1772); **CT arm;
     else lc_exit_days = (fup_days - 58.7590); **X-ray arm;
end;
else lc_exit_days = fup_days;  **deceased: no change;
```

	Actual PY through 12/31/09	Adjusted PY
Spiral CT	164483	164476
Chest X-ray	164626	164614
Total	329109	329090

Table 4: Actual and adjusted person-years of follow-up for a lung cancer incidence analysis with a cutoff of 12/31/09

The adjusted follow-up time is negative for alive individuals whose last contact was very early in the study; their time can be set to 0 without introducing much error (~ 17 person-years).

EVENT CONDITIONS

To identify events occurring through 12/31/09, use the following variables:

Event Type	**Condition**
Lung cancer diagnosis	CONFLC = 1
Lung cancer death	FINALDEATHLC = 1
Death from any cause	DEATHSTAT = 1, 2, or 3
Death from other specific cause	(DEATHSTAT = 1, 2, or 3) and (DCFICD = ICD-10 codes of interest)

Table 5: Event types and variables used to determine event status

To identify deaths occurring before 1/15/09, add the condition DEATHCUTOFF = 1 to the conditions listed above.

COMPARISON WITH PUBLISHED RESULTS

The following table compares the numbers published in the NLST primary results paper with those obtained by using the datasets as described in previous sections.

Event type (and cutoff date)	Statistic type	Study arm	Published numbers	Numbers from datasets
Lung Cancer Incidence (12/31/09)	Count	Spiral CT	1060	1089
Lung Cancer Incidence (12/31/09)	Count	Chest X-ray	941	969
Lung Cancer Incidence (12/31/09)	Rate*	Spiral CT	645	662
Lung Cancer Incidence (12/31/09)	Rate*	Chest X-ray	572	589
Lung Cancer Mortality (1/15/09)	Count	Spiral CT	356	359
Lung Cancer Mortality (1/15/09)	Count	Chest X-ray	443	448
Lung Cancer Mortality (1/15/09)	Rate*	Spiral CT	247	249
Lung Cancer Mortality (1/15/09)	Rate*	Chest X-ray	309	312
All-Cause Mortality (12/31/09)	Count	Spiral CT	1877	1904
All-Cause Mortality (12/31/09)	Count	Chest X-ray	2000	2025
All-Cause Mortality (12/31/09)	Rate*	Spiral CT	1121	1137
All-Cause Mortality (12/31/09)	Rate*	Chest X-ray	1202	1217

* per 100,000 person-years

Table 6: Comparison of published results with results obtainable using the datasets

FINAL NOTES

The participant dataset contains all variables necessary for the analyses described in this document.

If calendar dates are essential for a research project, a special request for additional date-related data may be made through CDAS. Such requests will be thoroughly reviewed and may be rejected.