
NLST DATA COMPARISON: QUERY TOOL, CDAS DATASETS

The National Lung Screening Trial (NLST) produced a collection of datasets and images for research. These datasets contain detailed information about participants in NLST, including demographics, screening exam results, diagnostic workup, cancer diagnosis, treatment, and mortality.

Investigators may obtain these NLST datasets from two different sources: the [CDAS](#) website and the TCIA Query Tool. Each source has a different emphasis, and some datasets are available from one source only. In order to receive the datasets from either source, investigators must first [submit a request](#) through the CDAS website and complete a data transfer agreement.

The 16 datasets on the CDAS website are zipped into a single file for easy downloading, and they contain comprehensive data on all NLST participants. The datasets are in SAS format, but other formats are available upon request. Images cannot be accessed through the CDAS website.

While the TCIA Query Tool contains the same comprehensive data found in the CDAS datasets, its emphasis is to enable users to identify subsets of the NLST population and get images and data on those sub-populations. It is a web-based application whose capabilities include:

- Search NLST database for sub-populations of interest
- Download data on sub-populations or the entire NLST population (in CSV format)
- Download CT images on sub-populations (in DICOM format)
- View pathology images (not downloadable)

CT and pathology Images may also be obtained on external hard drive; [contact CDAS staff](#) to coordinate the transfer. Pathology images are in Aperio SVS format.

DATA AVAILABILITY COMPARISON

The table below describes the differences in the datasets available through the two sources.

Sources	Dataset Name	Notes
Both (CDAS and Query Tool)	Participant dataset	CDAS's Participant dataset contains most of the useful NLST data. In the Query Tool, the Participant dataset has been split into 10 separate tables. All variables can be queried.
CDAS only	14 other main datasets	These data cannot be used to build queries in the Query Tool.
	LSS Health Assessment Questionnaire (HAQ) dataset	Not available in Query Tool
Query Tool only	"IMS Derived" variables	These variables were designed to facilitate easier queries and to make key information from some of the "14 other main datasets" available for querying.
	SCT Scanner data from DICOM headers ("SCTImageInfo")	Can be queried in Query Tool
	LSS pathology datasets	Can be queried in Query Tool

Table 1: Comparison of Data Availability from the two Sources