

A Query Tool for Investigator Access to the Data and Images of the National Lung Screening Trial

Paul K. Commean · Joshua M. Rathmell · Ken W. Clark ·
Dave R. Maffitt · Fred W. Prior

Published online: 5 March 2015
© Society for Imaging Informatics in Medicine 2015

Abstract The National Cancer Institute (NCI), in conjunction with *blinded university*, provides a mechanism to enable public access to the study data, CT radiology images, and pathology images from the National Lung Screening Trial (NLST). Access to the data and images is through the NCI-sponsored, blinded university-hosted The Cancer Imaging Archive (TCIA), a repository of more than 40 study collections of cancer images. Once access to the NLST data has been granted by NCI, a *Query Tool* within TCIA is used to access the NLST data and images. The Query Tool is a simple-to-use menu-driven database application designed to quickly pose queries and retrieve/save results (from 53,452 NLST participants), download CT images (~20 million available), and view pathology images (~1200 available). NLST study data are contained in 17 Query Tool tables with ~370 variables to query. This paper describes Query Tool design, functionality, and usefulness for researchers, clinicians, and software developers to query data, save query results, and download/view images.

Keywords Cancer imaging · Image archive · Database querying · Data repository · Clinical studies data access · NLST · Query tool · Pathology viewer · caMicroscope

P. K. Commean (✉) · K. W. Clark · D. R. Maffitt · F. W. Prior
Mallinckrodt Institute of Radiology, Washington University School
of Medicine, 510 South Kingshighway Blvd., St. Louis, MO 63110,
USA
e-mail: commeanp@mir.wustl.edu

J. M. Rathmell
Information Management Services, Inc., Rockville, MD, USA

Background

The National Lung Screening Trial (NLST) was a randomized controlled trial comparing the relative utility of chest radiographs and CT imaging for lung cancer screening. During 2002–2004, 53,452 participants aged 55–74, with a history of heavy smoking but no lung cancer, were randomly assigned to two study arms in equal proportions: low-dose spiral computed tomography (CT) or chest radiography (CR) [1]. Participants were offered three imaging screens at 1-year intervals. The NLST goal was to assess whether low-dose spiral CT screening reduces lung cancer mortality relative to CR in persons at high risk for developing lung cancer.

The NLST was a collaborative effort of the National Cancer Institute's Division of Cancer Prevention, which funded and administered the Lung Screening Study (NLST-LSS), and the American College of Radiology Imaging Network (ACRIN) [2], funded by the Division of Cancer Treatment and Diagnosis, which administered the NLST-ACRIN. Westat [3] collected and managed the NLST-LSS data, with statistical analysis provided by the Information Management Services (IMS) (Rockville, MD). NLST-LSS contracted *blinded university* to collect all LSS CT images; chest X-rays were retained by screening sites. ACRIN collected and managed the NLST-ACRIN data and radiology images, with statistical analysis provided by the Brown University (Providence, Rhode Island) ACRIN Biostatistics Center (Brown). Under harmonized protocols, the NLST was conducted through two networks of screening centers: NLST-LSS (10 centers; 34,612 participants) and the NLST-ACRIN (23 centers; 18,840 participants). Screening exams and image collection concluded in 2007 while follow-up data were collected into 2010. Trial results, announced in 2011, indicated that low-dose spiral CT screening reduces lung cancer mortality in persons at high risk for developing lung cancer and is superior to CR screening [4]. Following the conclusion of NLST, all non-

image data (hereafter, “data”) were delivered to IMS and all CT images to blinded university. The CT images have been added to The Cancer Imaging Archive (TCIA) [5, 6].

Access to NLST data, including images, has been restricted to approved cancer researchers by the National Cancer Institute. Under NCI Special Studies IRB protocol 13-C-N010 and the complementary blinded university IRB protocol, interested researchers must submit a brief summary of their proposed research and the data they require for NCI review and approval prior to being granted access to NLST data managed by TCIA. Proposals are filed and cleared through the Cancer Data Access System (CDAS) [7].

The NLST data and images constitute a wealth of cancer and mortality information from serial screens in a large narrowly defined population collected in a well-managed clinical trial [8]. The NLST data contain information collected using trial-approved LSS and ACRIN forms [9]. The form data were collected and initially processed by Westat (LSS) and by ACRIN Data Management staffs. IMS and Brown then processed the LSS and ACRIN form data into a harmonized data set for trial-wide analysis. IMS provided the harmonized data set to blinded university in a custom spreadsheet format. Blinded university organized these data in a PostgreSQL database, the Clinical Studies Database (CSDB). Blinded university also extracted scanner-event data from the DICOM [10] headers of the 73,000+ CT image exams archived at blinded university and added these metadata to the database. LSS pathology metadata were processed by Westat and IMS and then provided to blinded university, where they were also added to CSDB. The LSS pathology image collection includes ~1250 images of resected lung tissue from ~460 lung cancer patients in LSS.

Investigators can access the NLST data and images through the TCIA *Query Tool*, which provides an intuitive, user friendly interface for exploring NLST data. Query Tool is launched from the TCIA website [11] or from NCI’s Cancer Data Access System (CDAS) website [7] hosted by IMS. Alternatively, the harmonized data set is available (in SAS [12] format) from CDAS.

The Query Tool was developed to provide a simple user interface to permit researchers to access clinical studies information from the TCIA CSDB without the need to learn SQL. Query Tool also permits direct download of radiology images or viewing of pathology images that are identified in query results. As the user selects variables from the tables, the Query Tool automatically creates the appropriate SQL statement, which the user can execute by selecting the run tab to retrieve the corresponding results from the database and have them displayed. The Query Tool thin web client design goals were (1) to develop a simple user interface with no structured query language (SQL) knowledge required to run queries, (2) to enable easy setup and integration of a database into the Query Tool which allows multiple keys to create joins between

tables, (3) to allow easy selection of variables to be displayed in the results and limiting the results returned by using condition variables with drop-downs to make condition value selection, (4) to save the results of a query, (5) to download CT images or view pathology images based on the results returned from a query, (6) to implement the Query Tool to run on any computing platform, (7) to allow new features to be added to the Query Tool, and (8) to be the user interface for the NLST clinical trial data and images. This paper describes Query Tool design, functionality, and usefulness for researchers and clinicians, to query data, save their query results, download CT images, and view pathology images.

Methods

Data and Images

The clinical studies data from NLST are stored in a PostgreSQL database. PostgreSQL was chosen because it is a free, open-source application. The data consist of 17 tables (Table 1) that logically organize the information and are as consistent as possible with data that would be received from CDAS. The table names along with the Primary and Foreign Keys are listed in Table 1.

There are differences in the data available through the Query Tool and in the SAS formatted data files from CDAS. Both the Query Tool and the SAS data files from CDAS include the most commonly used data, which are contained in the “Participant” SAS file (Table 2). The “Participant” file consists of one record per participant in the NLST. The Query Tool represents the “Participant” data in Tables 1–8, 10, and 11. Data from other SAS files are not directly included in the Query Tool, but some are available in summarized form in the “IMS Derived” Tables 9 and 12 in the Query Tool. Summarized variables are more user friendly for building queries and interpreting results. For instance, the number of nodules seen on a participant’s last screening exam is contained in the Query Tool variable “lstscre_num_nodule”, which summarizes across multiple records from CDAS’s CT Abnormalities SAS file. Images and supporting image data are available only through the Query Tool such as the SCT Image Information Table 13 and the LSS Pathology Tables 14–17 which are not available through CDAS.

The “SCT Image Information” table was created from data obtained from the CT DICOM image headers for the participants participating in the CT arm of the trial. The Query Tool utilizes the series instance unique identifiers in conjunction with the TCIA image Download Manager to download the CT images resulting from a query.

The LSS pathology data (Tables 14–17) contain information obtained from a subset of LSS participants who were diagnosed with lung cancer and underwent surgical removal

Table 1 NLST Query Tool database tables

Tables	Primary keys ^a	Foreign keys ^a
1. Demographics	pid	pid
2. Lung cancer diagnosis	pid	pid
3. Smoking history	pid	pid
4. Death last contact EVP	pid	pid
5. Medical history	pid	pid
6. Family lung cancer history	pid	pid
7. Alcohol history	pid	pid
8. Work history	pid	pid
9. IMS-derived person variables	pid	pid
10. Screening results	pid, study_yr	pid, study_yr
11. Follow-up procedures	pid, study_yr	pid, study_yr
12. IMS-derived SCT screen variables	pid, study_yr	pid, study_yr
13. SCT image information	serial_number	pid, study_yr
14. LSS pathology tumor	pid, study_yr	pid, study_yr
15. LSS pathology donor block	pid, study_yr, tissue_block_id	pid, study_yr
16. LSS pathology region of interest	pid, study_yr, tissue_block_id, roi_number	pid, study_yr, tissue_block_id
17. LSS TMA core	pid, study_yr, tissue_block_id, roi_number, tmacore_id	pid, study_yr, tissue_block_id, roi_number

Spiral CT (SCT) is the legacy initialism used by the NLST because at the time of the study not all CT scanners were spiral

^a *pid* participant identifier, *study_yr* study year, *id* identifier, *roi* region of interest, *tma* tissue microarray

of lung tissue. Resected pieces of preserved lung tissue were obtained by NLST from pathology labs for the purpose of constructing tissue microarrays (TMAs) for later study. From each donor block of tissue, a thin slide was cut, stained, and digitally imaged (to guide TMA construction). These images and related pathology metadata are available through the Query Tool.

System Overview and Launching the Query Tool

The Query Tool application enables the user to pose queries that return results from the CSDB that contain text/numeric NLST data (Fig. 1). The amount of data and CT and pathology

images available to NLST Query Tool users is shown in Table 3. The text/numeric data can be saved to the user’s computer. CT images identified in query results can be downloaded using the TCIA Download Manager. Digitized pathology slide images can be viewed when the pathology viewer (caMicroscope [13, 14]) launches in the user’s web browser. The pathology viewer allows annotations to be performed on the images such as creating rectangles, circles, polygon, or free hand drawing with the ability to add a description to each type of drawing.

The Query Tool application may be launched from (1) TCIA [4] (cancerimagingarchive.net) and (2) CDAS [7] (biometry.nci.nih.gov/cdas). Each website requires the user to have obtained permission to access the NLST data and be logged in to enable launching the Query Tool. If the user logs into TCIA, a “Search NLST Data” button appears, and if the user logs into CDAS, the “Launch the TCIA Query Tool” link is provided.

Table 2 NLST data availability from Query Tool and CDAS SAS files

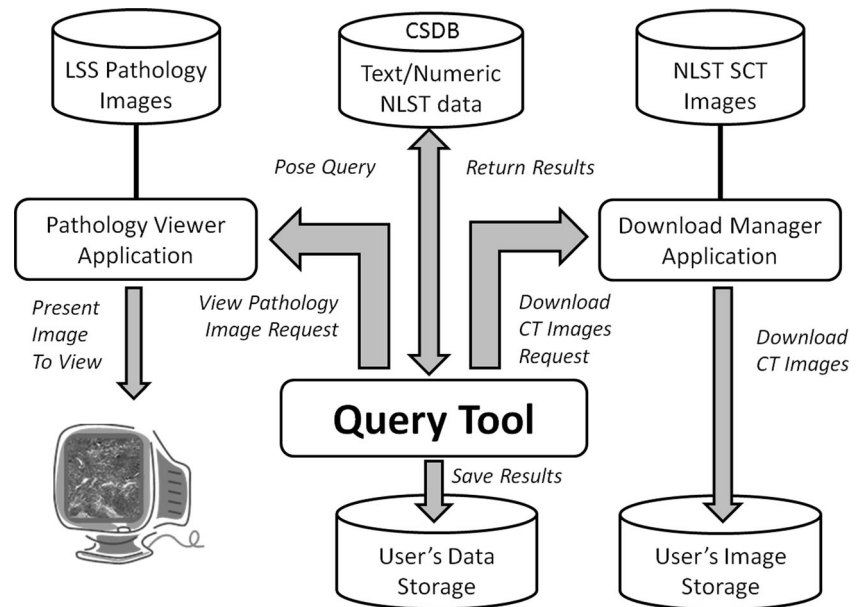
Query Tool tables	CDAS SAS filenames ^a
Tables 1–8, 10, 11	Participant
Tables 9 and 12	CT abnormalities, CT comparison read abnormalities participant
Table 13	No data available
Tables 14–17	No data available
No data available	CT screening; chest X-ray screening Chest X-ray abnormalities Chest X-ray comparison read abnormalities Medical complications; treatment Lung cancer progression; LSS non-cancer conditions ACRIN non-cancer conditions; LSS HAQ

^a See <https://biometry.nci.nih.gov/cdas/nlst/datasets/comprehensive/>

Database Design

The Query Tool application was developed to allow easy extension to data sources other than those from NLST. Several additional tables have been added to the CSDB to support the display of the Query Tool graphical user interface (GUI). The additional tables are (1) table names containing the names of each of the CSDB data tables with the order of display and whether the table is displayable in the graphical user interface; (2) table variable names containing the table name, variable name, data type (text, integer, real), displayable, tool tip

Fig. 1 System overview—The Query Tool enables the user to create queries and return results from the NLST data contained in the Clinical Studies database (CSDB). The results can be saved, or used to download the NLST Spiral CT (SCT) images, or view the LSS Pathology images



descriptions for each variable, and display format columns; (3) table variable values containing the table name, variable name, variable values, and drop down ordering of the values within the drop down box; (4) a table containing the primary keys, and a table containing the foreign keys, respectively, for each of the CSDB data tables; (5) a table to keep track of the users who access the Query Tool with date and time; and (6) a table to record each image series download attempt with user information, date, and time. With the addition of the graphical user interface tables, only a few variables, document names, URLs, and other minor items need to be changed to accommodate data from any other clinical study within CSDB and the Query Tool's source code.

Query Tool Features

The Query Tool's graphical user interface gives the user the ability to select variables from the 17 tables shown in the "Select Returned Values" tab (Fig. 2) by clicking on each variable needed in the query (Fig. 3). Constraints can be added to variables by choosing limit values for each condition variable found in the "Add Constraints" tab. The select return values and constraint variables are placed into the table column named "Selected" (Fig. 3). The Query Tool automatically creates the

Table 3 Amounts of data available to Query Tool users

Type of data	Amount
Text/numeric metadata	17 tables, 400+ variables, 270+ megabytes
Radiology CT images (DICOM) [10]	20+ million, ~10.5 terabytes
Pathology images (Big TIFF) [15]	1200+, ~1.2 terabytes

SQL statement as the user selects variables and includes constraints. Once a query is created, the query can be run or modified to create a new query to run. After a query is run, the results are automatically displayed in the Run Query/View Results tab, and the results can be saved. If the user wishes to download associated CT DICOM images for the participants who were enrolled in the spiral CT arm of the study, the "Download Associated CT Images" button allows immediate downloading using the TCIA Download Manager window.

The two menu bar items allow Query Tool users to view publications and obtain help. Publications relevant to the NLST study can be viewed/downloaded. The Help menu includes the Query Tool User's Manual, Data Dictionaries, Tutorial, NLST Manual of Operations and Procedures, Study Forms, NLST TCIA wiki, other important websites, and contact information.

To create a query, a user would select the triangle in front of a table listed under the "Available" column in the "Select Returned Values" tab. Demographics table and the Demographics folder will open to display the columns as shown in Fig. 3 that can be selected for display in the "Run Query/View Results" tab when the tab is selected. When the mouse arrow is hovered over the Demographics variable named "rndgroup" a tool tip appears to give more details about the variable. After selecting the "rndgroup" variable with the left mouse button, the Demographics.rndgroup variable appears in the "Selected" column (Fig. 3). Fig. 4a shows selections from multiple tables (demographics, smoking history, and SCT image information), and Fig. 4b shows a variety of constraints added to limit the returned results (weight between 160 and 200, smoking history of 50 pack-years and fewer than 30 cigarettes a day). The results can be seen in a scrollable "Run Query/View Results" window (Fig. 4c) containing the selected return values limited by the constraints (Fig. 4b). The

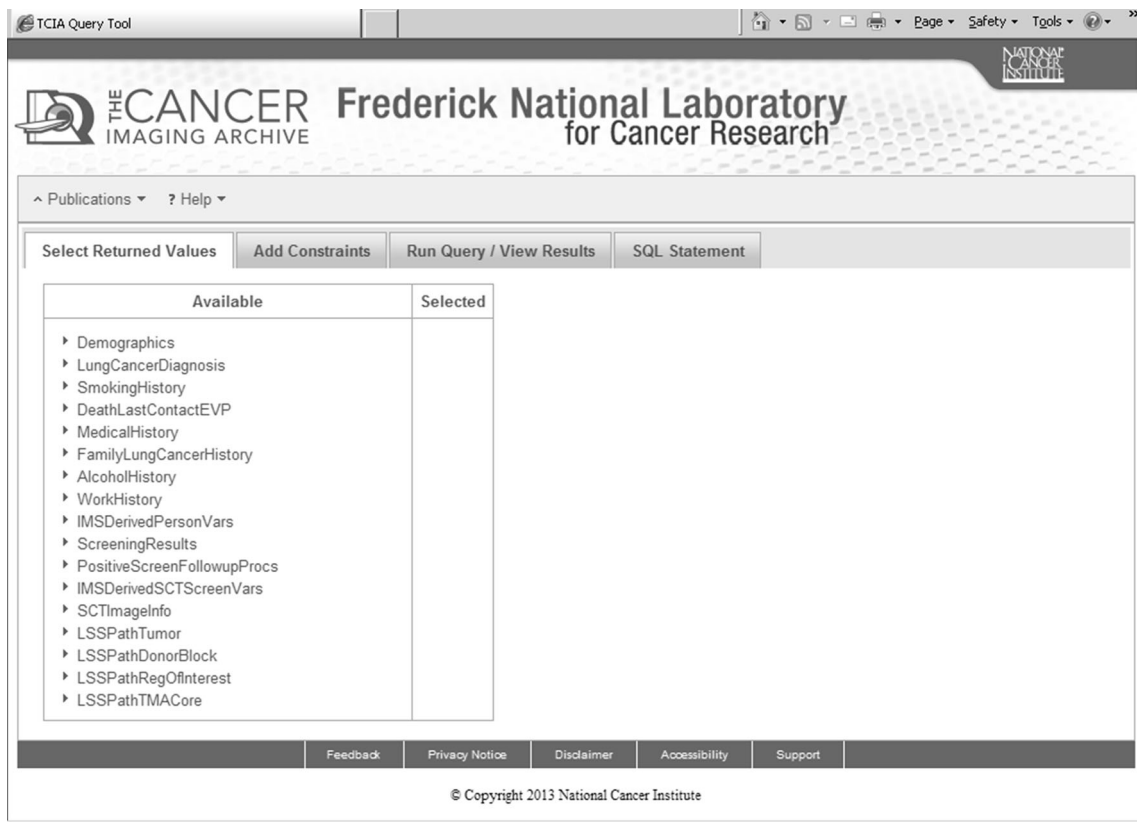


Fig. 2 NLST Database Query Tool graphical user interface has a menu bar to enable viewing of publications and provides documents/links to help with using the Query Tool. The four tabs allow selecting variables from the 17 tables to be returned to view the results when the run query

tab is selected. Constraints can be added to the query to specify the type of results returned. The SQL statement is created as variables are selected and constraints added

results can be sorted by selecting either the up or down arrow to the right of each table.variable name. The Number of Records is shown (i.e., 1409) and results are presented in a paginated format (i.e. 29 pages with 50 records or rows per page).

The SCT DICOM images collected during the NLST are available for downloading after creating and running a query. Select the “Download Associated CT Images” button located in the “Run Query/View Results” tab, and the Download Manager window will appear (Fig. 5). Selecting the “Start” button starts the download process with the status being given regarding the progress of the download, such as Complete, Downloading, or Not Started (Fig. 5).

The LSS pathology images from resected lung tissue for the ~460 lung cancer patients in the LSS portion of the NLST study can be viewed by selecting the variable `pathology_image` from the `LSSPathDonorBlock` table to the query. After running the query, a `ViewPathologyImage` button will appear for those participants that had lung cancer in the study, and a tissue block was resected with `pathology images` created. Selecting the `ViewPathologyImage` button opens a pathology image viewer, `caMicroscope` [13, 14] as shown in Fig. 6. The pathology viewer allows annotations to be drawn on the image and saved for later use.

Fig. 3 The “Select Returned Values” tab can be utilized to create a query. The mouse arrow can be used to hover over a variable to highlight a variable and display the tool tip. In this case, `Demographics.rmdgroup` was highlighted and selected to appear in the “Selected” column



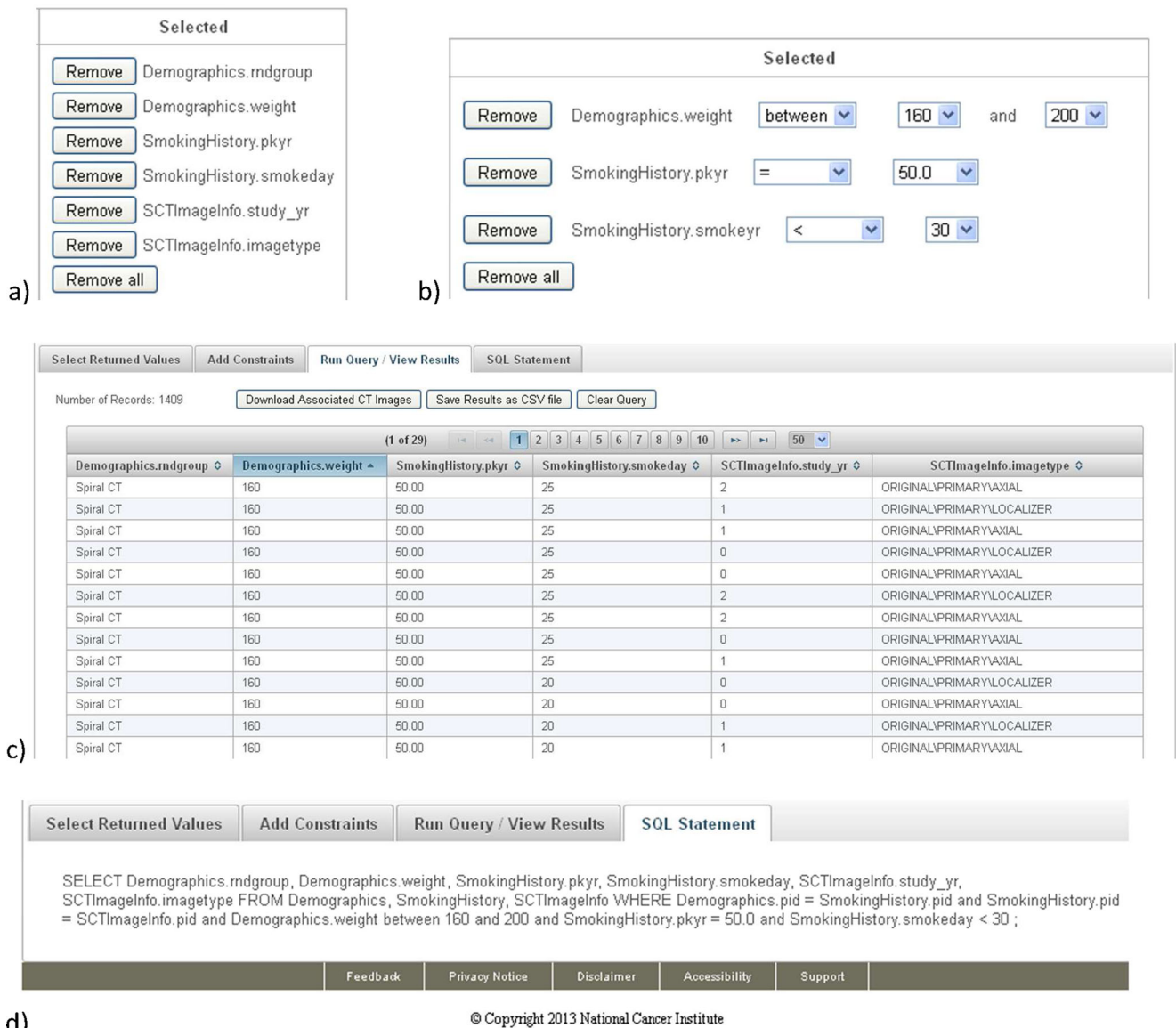


Fig. 4 An example query with **a** selected variables for display in the Run Query tab, **b** constraint variables with limit values chosen, **c** results from running the query, and **d** SQL statement created by Query Tool

Fig. 5 Download Manager used to download the NLST SCT DICOM images. The directory location for downloading the files can be selected

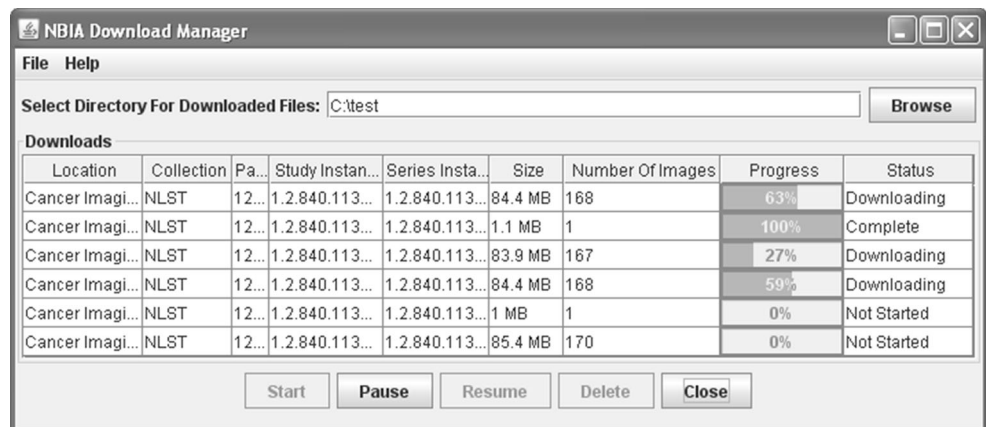


Fig. 6 The pathology image is a single slice from a tissue donor block for a participant with lung cancer

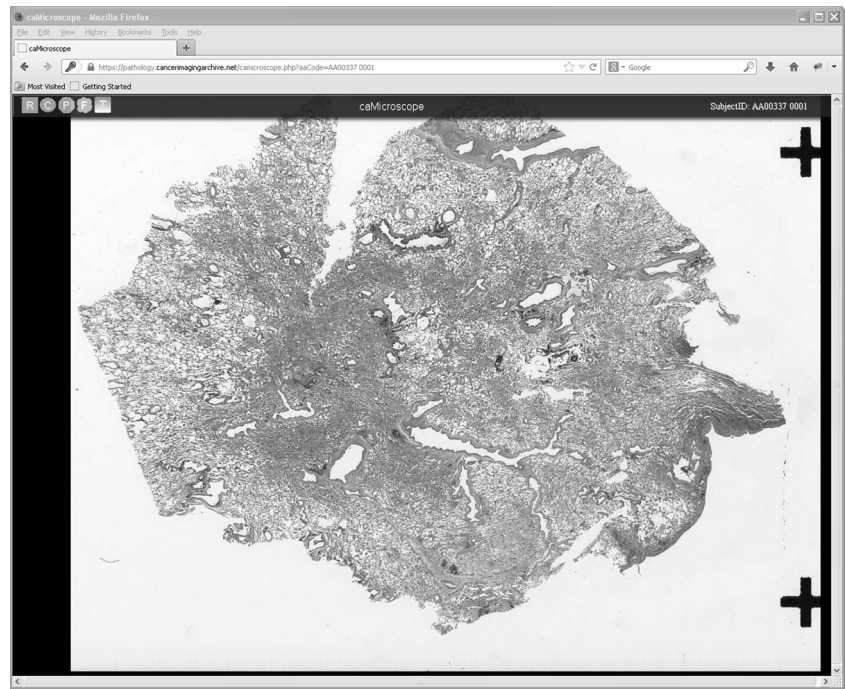


Table 4 Query tool run times for different complexities of queries

Query Number	Selected variables	Condition variables	Condition values	Results returned	WashU runtime	IMS runtime
1	Demographics.mdgroup Demographics.weight SmokingHistory.pkyr SmokingHistory.smokeday SCTImageInfo.study_yr SCTImageInfo.imagetype	Demographics.weight SmokingHistory.pkyr SmokingHistory.smokeday	Between 160 and 200 =50 <30	1409	1 s	1 s
2	Same variables as query 1	Demographics.weight SmokingHistory.pkyr SmokingHistory.smokeday SCTImageInfo.study_yr SCTImageInfo.imagetype	Between 160 and 200 =50 <30 =2 =originalprimaryaxialhelix	9	1 s	1 s
3	Same variables as query 1	No conditions selected	Not applicable	201,637	3 s	6 s
4	Demographics.mdgroup Demographics.weight SmokingHistory.pkyr SmokingHistory.smokeday	No conditions selected	Not applicable	53,452	1 s	1 s
5	Demographics.mdgroup Demographics.weight SmokingHistory.pkyr SmokingHistory.smokeday SCTImageInfo.study_yr SCTImageInfo.kvp	Demographics.weight SmokingHistory.pkyr SmokingHistory.smokeday SCTImageInfo.study_yr SCTImageInfo.kvp	Between 170 and 180 =50 <30 =0 =80.00000	8	1 s	1 s
6	Demographics—all AlcoholHistory—all SCTImageInfo—all	Demographics.age Demographics.gender	=70 =Female	1768	1 s	1 s
7	Demographics—all AlcoholHistory—all SCTImageInfo—all LSSPathTumor—all LSSPathDonorBlock—all LSSPathRegOfInterest—all LSSPathTMACore—all	No conditions selected	Not applicable	12,802	6 s	6 s

Note: Run times and CT image download times may vary depending on the network speed, server loading, and your computer’s capabilities

Results

The Query Tool was evaluated to determine the run times for different complexities of queries at two different sites (Table 4). Blinded university, where the Query Tool was developed and where the NLST data and images reside, was one test site. The second site was IMS located in Rockville, MD. Running Query 1 (shown in Fig. 4) took about 1 s or less with 1409 results returned. For Query 2, two additional condition variables were added with only 9 results returned with a similar run time. Query 3 is the same as Query 1 with no conditions selected and took approximately 3 s to run returning 201, 637 results. For Query 4, removing the two SCTImageInfo variables from Query 1 with no conditions selected took about 1 s or less to return 53,452 results. Running Query 5 resulted in only 8 spiral CT images to be returned. By selecting the “Download Associated CT Images” button located in the “Run Query/View Results” tab, the NBIA Download Manager contained 11 images with a download time of approximately 3 s and approximately 7 Mbytes of image data. Queries 6 and 7 were created to test more complex queries where all of the variables were chosen from the tables. Query 2 was used to test a larger “Download Associated CT Images” run time for the nine returned axial helix series instances. These nine series contained 1420 images with a total size of approximately 722 Mbytes and a download time of approximately 2 min and 40 s from blinded university.

Discussion

Before the Query Tool, if an investigator wanted to query the NLST database for a research project, several labor-intensive steps were required. To subset the data to the population of interest for answering a research question, programming had to be done either by the investigator’s research team or by NLST support programmers. If images were requested, a list of images had to be communicated to NLST support staff and then delivered to *blinded university* where the images were extracted from the archive and written to electronic media that were then shipped to the investigator. The tasks of sub-setting the data, communicating between groups, and extracting images resulted in a long turn-around time between the question(s) posed and the images delivered. The Query Tool solves these issues by allowing the investigator to automatically create an SQL statement, run the query, refine the results based on the data returned, save the results to a file amenable to statistical analysis, and download the images immediately if desired.

Currently, since the NLST trial data has restricted access, the Query Tool is only available to approved users who (1) agree to data use policies, (2) complete a project application, (3) indicate the type of request, and (4) list all users who you

approve to access the data. Permission can be obtained through the CDAS website [7]. The Cancer Imaging Archive and CDAS websites, from which the Query Tool can be launched, require the user to log into the website and have the correct credentials before running the Query Tool.

While the Query Tool is implemented exclusively for TCIA-NLST, the Query Tool can be extended such that additional data repositories could be integrated into its user interface. The Query Tool allows easy database setup for a new clinical data repository, and is easily extensible. The Query Tool can be extended to allow the database to dynamically grow with automatic or semi-automatic data addition for ongoing clinical trials, studies, or projects.

There are several limitations associated with the current Query Tool

- The Query Tool does not incorporate a radiology viewer to display images, but a viewer is planned for a later release.
- The Query Tool does not have a method for automatically importing a new database into its graphical user interface.
- The Query Tool does not contain any features to create reports either manually or automatically.

Conclusion

The goals for designing the Query Tool were achieved, and the Query Tool is now available for researchers and clinicians. The wealth of NLST data and images available to investigators is likely to inform how and when medical practitioners use CT image screening for their patients at high risk for lung cancer. The Query Tool is designed to help guide and facilitate how investigators mine these data and images. The Query Tool would empower researchers to share their research data with others, enabling open science goals such as providing public access to scientific data via a web-based tool to allow scientific collaboration.

References

1. Aberle DR, et al: The National Lung Screening Trial: overview and study design. *Radiology* 258(1):243–53, 2011
2. ACRIN. *American College of Radiology Imaging Network*. [cited 2014 3/21/2014]; Available from: <http://www.acrin.org/HOME.aspx>.
3. WESTAT. *Westat*. [cited 2014 3/21/2014]; Available from: <https://www.westat.com>.
4. Aberle DR, et al: Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 365(5):395–409, 2011
5. Prior, F.W., et al. TCIA: An information resource to enable open science. *Engineering in Medicine and Biology Society (EMBC)*,

- 2013 35th Annual International Conference of the IEEE. IEEE, Osaka, 2013, pp 1282–1285
6. Clark, K., et al., The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging* 26(6):1045–1057, 2013
 7. NCI. *National Cancer Institute Cancer Data Access System (CDAS)*. [cited 2013 10/31/2013]; Available from: <https://biometry.nci.nih.gov/cdas/nlst/requests/instructions>.
 8. National Lung Screening Trial Research Team: The national lung screening trial: overview and study design. *Radiology* 28,2010
 9. NCI. *National Cancer Institute NLST Forms*. [cited 2013 10/31/2013]; Available from: <https://biometry.nci.nih.gov/cdas/nlst/forms>.
 10. NEMA. *Digital Imaging and Communications in Medicine (DICOM)*. [cited 2013 10/31/2013]; Available from: <http://medical.nema.org>.
 11. Prior, F.W., et al. The Cancer Imaging Archive. 2012 [cited 2013 10/31/2013]; Available from: <http://www.cancerimagingarchive.net>.
 12. SAS. *Statistical Analysis Software*. [cited 2013 10/31/2013]; Available from: <http://www.sas.com>.
 13. Wang, F., Jun Kong, Lee Cooper, Tony Pan, Tahsin Kurc, Wenjin Chen, Ashish Sharma et al., *A data model and database for high-resolution pathology analytical image informatics*. *Journal of pathology informatics*, 2011. 2(1): p. 32.
 14. *caMicroscope*. 2009 [cited 2013 10/31/2013]; Available from: <http://imaging.cci.emory.edu/wiki/display/CAMIC/Home>.
 15. *BigTIFF*. [cited 2012 10/31/2013]; Available from: <http://bigtiff.org>.