J Digit Imaging (2015) 28:18–23
DOI 10.1007/s10278-014-9720-1

# Data-Driven Decision Support for Radiologists: Re-using the National Lung Screening Trial Dataset for Pulmonary Nodule Management

**James J. Morrison · Jason Hostetter · Kenneth Wang · Eliot L. Siegel**

Published online: 26 June 2014
© Society for Imaging Informatics in Medicine 2014

**Abstract** Real-time mining of large research trial datasets enables development of case-based clinical decision support tools. Several applicable research datasets exist including the National Lung Screening Trial (NLST), a dataset unparalleled in size and scope for studying population-based lung cancer screening. Using these data, a clinical decision support tool was developed which matches patient demographics and lung nodule characteristics to a cohort of similar patients. The NLST dataset was converted into Structured Query Language (SQL) tables hosted on a web server, and a web-based JavaScript application was developed which performs real-time queries. JavaScript is used for both the server-side and client-side language, allowing for rapid development of a robust client interface and server-side data layer. Real-time data mining of user-specified patient cohorts achieved a rapid return of cohort cancer statistics and lung nodule distribution information. This system demonstrates the potential of individualized real-time data mining using large high-quality clinical trial datasets to drive evidence-based clinical decision-making.

**Keywords** Decision support · Data mining · Decision support techniques · Web technology

## Introduction

Cancers of the lung and bronchus have the highest age adjusted mortality of all cancers in the USA. In 2013, there were an estimated 159,480 deaths from lung cancer, accounting for almost a third of all cancer-related mortality [1]. Prognosis is poor, with a survival rate of only 16 % at 5 years [2]. With a high relative incidence and poor outcome, lung cancer presents a large target for cancer prevention, diagnosis, and treatment efforts.

As with all cancer, early detection and intervention are crucial to improving long-term outcomes. Indeterminate pulmonary nodules have long posed a diagnostic dilemma, particularly in low-risk patients, as a considerable overlap exists in the imaging appearance of both benign and malignant etiologies. Recent research efforts such as the National Lung Screening Trial (NLST) have focused on the efficacy of lung cancer screening programs in high-risk populations. Enrolling 53,451 asymptomatic patients, the NLST demonstrated a relative reduction in lung cancer mortality of 20 % in patients between 55 and 74 years of age with at least 30 pack-years of smoking history who were screened using low-dose CT [3]. In response to this evidence, the US Preventative Services Task Force released new recommendations in favor of annual low-dose CT screening of high-risk patients based upon age and smoking history [4]. Despite the recommendation, the combined cost of imaging, diagnostic, and surgical interventions required to diagnose and treat a single cancer presents a challenge to the widespread implementation of screening programs [5]. Intelligent decision support tools go beyond a one-size-fits-all approach to pulmonary nodules by leveraging additional patient and nodule characteristics from a large cohort clinical dataset such as the NLST to tailor the recommendations for each patient. These tools can help mitigate diagnostic and treatment costs through identifying the most appropriate screening population, optimizing follow-up imaging and intervention, and increasing adherence to evidence-based practice.

Evidence-based medicine is exemplified by clinical decision-making that integrates clinical experience with best

J. J. Morrison (✉) · J. Hostetter · E. L. Siegel
Department of Radiology, University of Maryland,
22 S. Greene St., Baltimore, MD 21201, USA
e-mail: jjmorrison@gmail.com

K. Wang · E. L. Siegel
Baltimore VAMC, Baltimore, MD 21201, USA

available scientific literature, clinical trial results, and data analysis [6]. Real-time clinical decision support tools can incorporate the latest validated research and resources to individualize screening, diagnosis, and treatment of lung cancer while optimizing utilization of limited healthcare resources such as CT. Several types of artificial intelligence techniques have been used in clinical decision support systems including rule-based reasoning, Bayesian networks, neural networks, and case-based approaches. Previous research has shown that clinical decision support tools are accurate in qualitative diagnosis and forecasting malignant probability of pulmonary nodules [7]. After lung cancer diagnosis, Sesen et al. demonstrated the use of a Bayesian network decision support system for personalized survival estimates and treatment selection recommendations in lung cancer [8]. While not directly comparable to lung nodule screening, evidence-based clinical decision support tools have also been effective in decreasing the overutilization of CT while increasing yield of the studies performed for CT pulmonary angiography [9].

We developed a clinical decision support application that allows users to define a clinical case and find matching pulmonary nodules in NLST patients. The follow-up interval and outcome information from the matched cohort may then be used to tailor individual patient management. Real-time data mining of high-quality clinical research datasets like the NLST enables a more sophisticated approach to evaluating individual risk and determining follow-up for a given pulmonary nodule.

## Materials and Methods

The NLST dataset is available through the Cancer Data Access System (CDAS) run by the National Cancer Institute/National Institutes of Health. After registration with CDAS, a project proposal was submitted, approved, and a data transfer agreement was signed between the authors and the National Cancer Institute (NCI). The standard agreement permits access to the dataset for use solely in the proposed research plan, including access to the chest radiograph and low-dose computed tomography (LDCT) images. Personally identifiable health information was not provided for any patient in accordance with the Health Information Portability and Accountability Act (HIPAA), and researchers are restricted from attempting to identify participants. The NCI retains ownership of the data, and an acknowledgment is required in each publication resulting from the use of the data as well as the submission of a description of each publication on the CDAS website. The data transfer agreement allows access for 3 years, beyond which a new agreement or amendment is required.

Four tables from the dataset provide the source material for the nodule queries. Specifically, a "Participant" data table contains a unique identifier for each patient along with all associated patient demographic information. The second table, describing the "Spiral CT Abnormalities," provides information on every pulmonary nodule identified on spiral CT for each screening year. For nodules greater than 4 mm, the data table includes nodule size (longitudinal diameter and longest perpendicular diameter) and descriptors of margins, opacity, and lobe location. Additional information on follow-up recommendations and nodule changes between screening examinations are available in the "Spiral CT Comparison Read Abnormalities" and "Spiral CT Screening" tables.

A cloud-based virtual private server was created from an internet hosting service (DigitalOcean, Inc., NY, USA) containing a 20 GB solid state drive, 512 MB memory, and Ubuntu 12.04 LTS Linux to host the application [10, 11]. The application was produced using Node.js, a highly scalable software platform supporting both server and client-side JavaScript development; MySQL, a widely used open-source relational database management system (RDBMS) loaded with the NLST data; and NGINX, an open-source HTTP server used to route client traffic to our application [12–14]. JavaScript was chosen as the development language for its cross-platform compatibility, ease of rapid development, and minimum of overhead. Communication between Node.js and MySQL was possible through a third-party plug-in library permitting dynamic queries of the NLST data tables [15].

The browser client and user interface were developed using a combination of Bootstrap, AngularJS, and jQuery. Bootstrap is an open-source collection of HTML and cascading style sheet (CSS) tools for rapidly producing an organized, visually clean, and responsive user interface [16]. This ensures that the application interface is usable and consistent across different web browsers and user platforms (e.g., PC, Mac, Tablet, Smartphone). AngularJS provides a framework that automates binding and synchronization of data elements between the HTML5 user controls and the client-side JavaScript data processing [17]. The jQuery library was used to further simplify the client-side script building through the use of several sub-components: jQuery UI for user interface elements such as slider bars, jQuery UI Touch Punch for compatibility with gesture-based devices such as smartphones and tablets, and jqPlot for graphical output [18]. The software components were chosen for their wide availability, extensive developer community support, and free- and open-source (FOSS) licensing models. Testing was performed using the following web browsers: Chrome 31.0.1650.63 (Google, Inc. CA, USA), Safari 7.0.1 (Apple, Inc. CA, USA), and Safari for iOS 6 and 7 (Apple, Inc., CA, USA).

The client-side interface is divided into a search builder and a result display. The search builder consists of range-select sliders, radio buttons, hyperlinks, and checkboxes allowing a user to select patient characteristics and nodule descriptors to reflect a clinical case. Specifically, these input elements include patient age, gender, smoking history in years or pack-

years, nodule size, nodule density, nodule margins, and lobe location (Fig. 1). Once the user input is complete, a search is initiated by clicking the "Find Nodules"' button beneath the search builder interface. A client-side data array containing each of the user selections is then sent asynchronously to the Representational State Transfer (REST) Application Programming Interface (API) exposed by the server.

The server-side application listens for client search requests. When a search request is received, the user-defined data array is parsed, and a Structured Query Language (SQL) command is built specifying which patients and nodules to select from the data tables. Utilizing the efficiency of RDBMS data processing techniques, a single SQL query is able to unite the four data tables and extract the desired information. The query results are returned in a JavaScript Object Notation (JSON) format enabling simplified server-side data processing. Each element in the returned data structure represents a unique nodule and contains a patient identifier, year in which lung cancer was diagnosed, longest nodule diameter, and nodule lobe location. Additional server-side data processing is performed to produce the descriptive statistics including number of nodules matching the specified criteria, number of lung cancers associated with the selected nodule cohort, mean time to lung cancer diagnosis (of those nodules found to be malignant during the study period), distribution of matching nodules by pulmonary lobe, and Fleischner recommendations by size categories. The processed data is then returned to the client and displayed on screen in textual, tabular, and chart formats (Fig. 2).

During development, our initial search query response times were on the order of seconds to minutes, resulting in a sluggish user experience. This was found to be the result of performing mathematical set operations on the non-indexed SQL tables. Optimizing SQL table operations was achieved by defining a unique identifier in the data tables, known as a primary key index (in this case the patient identification number). We also tested a nested SQL query syntax where a child query is performed first, and the larger parent query is performed on the result of the child query.

Search return time performance using the described queries was tested by automating 100 queries against the cloud server hosting our data. Input parameters were randomized within the ranges specified by our user interface to simulate the potential query conditions. These randomized queries were performed using the nested and non-nested SQL query syntaxes, with and without a primary key index.

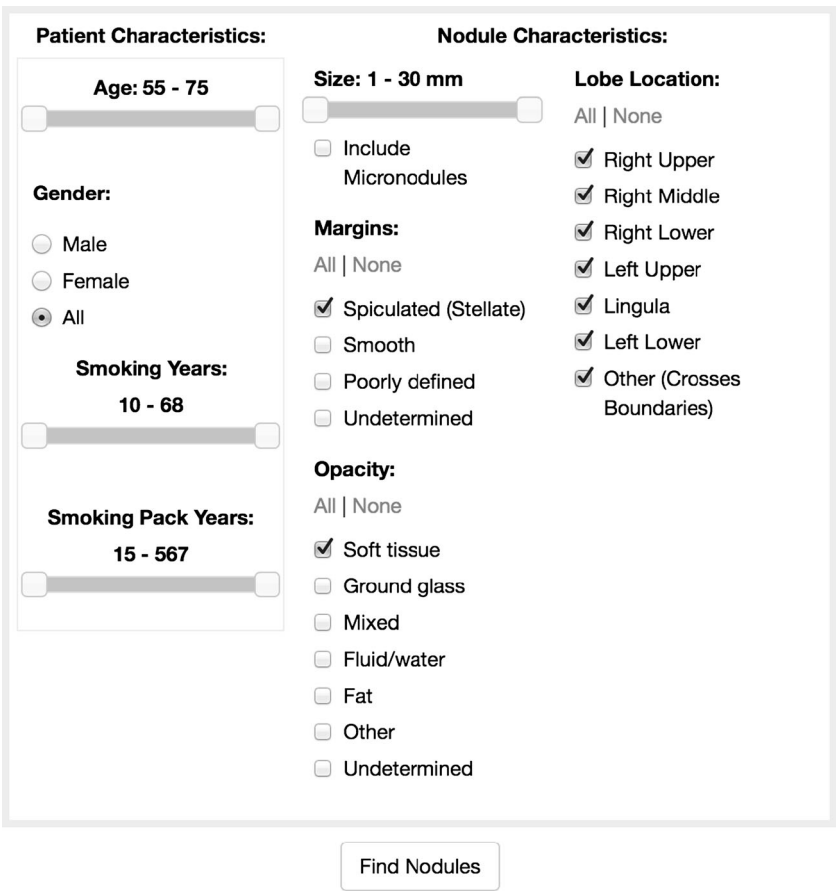


**Fig. 1** Using the query builder interface above, users can select a patient cohort by specifying a range of patient and nodule characteristics to search through the NLST data tables. Once the variables have been chosen, the user then clicks the "Find Nodules" button to execute the query


**Fig. 2** The result display is generated in real time from the query results determined by the input parameters selected in the query builder

## Results

Through the described interface, a user can interact with data from 26,721 patients randomized to the spiral CT screening arm of the NLST. A total of 58,589 pulmonary nodules are searchable: 33,816 nodules measuring greater than or equal to 4 mm, and 24,773 micronodules measuring less than 4 mm. The nodules measuring greater than or equal to 4 mm in the greatest dimension contain information on lobe location, margins, and opacity. Descriptive data was not recorded for micronodules; however, information on patient lung cancer outcome was documented. Confirmed lung cancers were diagnosed in 3,240 patients (12 %).

Depending on user selections, the application performs searches using all or a subset of four participant attributes (with 155,287 unique combinations) and four nodule attributes (with 301 unique combinations) combining for a total of $7.3 \times 10^{18}$ query possibilities. The open-ended nature of the search input presented a challenge to real-time data mining as result return times vary depending on the number and range of criteria entered.

The non-nested syntax without a primary key index provided the slowest response times with a mean of 2.8 s (St. Dev. 3.7 s, range 0.41–21.2 s). This was slightly improved by using the nested syntax, which gave a mean response time of 1.6 s (St Dev. 3.2 s, range 0.4–24.3 s). After adding a primary key index, response times improved markedly with the non-nested syntax, with a mean of 0.4 s (St. Dev. 0.1 s, range 0.2–1.3 s). Mean response times did not change for the nested syntax after adding the primary key index (1.6 s); however, the

standard deviation did decrease (3.2 s without index; 2.2 s with index) (Table 1).

Initially, a single search request generated multiple SQL queries, one for each type of output. For further simplification and performance optimization, these individual queries were refactored into a single query, thereby removing unnecessary overhead and increasing overall response times. The application currently uses a single query with non-nested syntax, as its linear query logic is simpler to understand and implement while providing the best performance after the addition of a primary key index. Using this optimized, non-nested, indexed query structure, 100 randomized queries were performed resulting in a mean query time of 0.3 s (St. Dev. 0.06 s, range 0.3–0.5), an overall improvement of 89 % in mean query response time compared to our original non-nested, non-indexed multiple query implementation.

## Discussion

The system described above represents a case-based reasoning approach to clinical decision support utilizing the NLST data as the pre-defined set of example cases [19]. The effectiveness of similar case-based clinical decision support systems has previously been shown for both vertebral compression fractures and breast MRI using smaller collections of comparison cases and example images to assist in prognosis and treatment [20, 21].

To the best of our knowledge, this is the first case-based decision support system to use a dataset repurposed from a large clinical research trial to tailor results to individual patient and disease-specific factors. The end goal of most clinical research trials is to affect best practices and patient care, usually through scientific paper publications. Large, high-quality clinical datasets have the added potential to inform clinical decision support systems. Additional sources of large, high-quality data exist, opening the potential for clinical applications similar to one the described above. Through this model, research data becomes directly applicable to practitioners and their patients, making available personalized, evidence-based diagnostic recommendations.

**Table 1** Query response times by SQL syntax

| Number of queries | Nested/non-nested | Primary key index | Query time (seconds) |
|---|---|---|---|
| Multiple | Non | No | 2.8±3.7 |
| Multiple | Non-nested | No | 1.6±3.2 |
| Multiple | Non | Yes | 0.4±0.1 |
| Multiple | Non-nested | Yes | 1.6±2.2 |
| Single | Non-nested | Yes | 0.3±0.06 |

There are several limitations to the current application. In general, data mining is constrained by the type and amount of data contained in a dataset. The NLST dataset contains 16 tables of data of which the described application only uses a subset—the 4 tables describing the pulmonary nodules on CT. Within the four data tables that are searched, we are again using only a subset of available data. There are 6 nodule attributes and almost 200 participant attributes (including personal demographics, work history, medical and family histories, etc.). Utilizing these additional data points and tables could permit further individualization of the descriptive queries; however, the specificity comes at a cost. As the search parameters become more personalized, the size of the returned patient cohorts can become smaller, which may present misleading data.

Additional limitations are inherent to the dataset. The NLST tracks the development of lung cancer by lobe. When multiple nodules are present in a lobe that eventually develops cancer, each one of those nodules are assigned a cancer prognosis. A cancer diagnosis cannot be directly matched to a specific nodule of origin. Similarly, individual nodules cannot easily be followed longitudinally through the years of the study, particularly when multiple nodules exist in the same lobe. Each nodule is given an identifier that is only unique for a given year. The following year, if the nodule persists, it is again given a unique identifier, but not necessarily the same identifier as the year prior. As such, the data is restricted to describing change between single study years. The source imaging studies are available making it possible to retrospectively assess all the described nodules and address these deficiencies.

Future directions for this work include the development of an on-the-fly regression model or machine learning algorithm to overcome the limitations of smaller cohort samples with highly specific queries. In its current state, the application simply displays a descriptive analysis of the NLST data. The application of additional statistical analysis would provide additional confidence information about the query results. Currently, unused data within the remaining tables contains information on the staging of diagnosed lung cancers and the interventions performed. Descriptive and statistical analysis based on this more detailed information may enhance prognostic and clinical decision support capabilities. Finally, the server-side application utilizes a REST API that allows for easy integration into applications other than our web-based client interface. Optimally, the query search and return could be integrated into a PACS client or reporting system, avoiding the need to enter the patient and nodule data manually.

## Conclusion

Clinical decision support systems are poised to play a central role in evidence-based medical practice and meaningful use of imaging resources. Our application repurposes data from a large, high-quality clinical study to form the core of a clinical decision support engine. Utilization of publicly available data and open-source technologies lowers the barriers to development of such tools and promotes patient care through the re-use of research data.

**Acknowledgments** The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

## References

1. American Cancer Society. Cancer Facts & Figures 2013. Available from: http://www.cancer.org/research/cancerfactsfigures/cancerfactsfigures/cancer-facts-figures 2013. Accessed 14 February 2014.
2. National Cancer Institute. Surveillance, Epidemiology, and End-Results Program. Available from: http://seer.cancer.gov/statfacts/html/lungb.html. Accessed 25 December 2013.
3. Aberle DR, DeMello S, Berg CD, Black WC, Brewer B, Church TR, et al: Results of the two incidence screenings in the national lung screening trial. N Engl J Med 369:920–31, 2013
4. United States Preventative Services Task Force. Lung Cancer Screening 2013 Available from: http://www.uspreventiveservicestaskforce.org/uspstf/uspslung.htm. Accessed 19 December 2013.
5. Tota JE, Ramanakumar AV, Franco EL. Lung cancer screening: review and performance comparison under different risk scenarios. Lung 192:55–63, 2014
6. Reiner BI, Siegel EL: The clinical imperative of medical imaging informatics. J Digit Imaging 22:345–7, 2009
7. Chen W, Liu J, Chen Q, Li W, Xiong Z, Long X: Bayes analysis in clinical decision-making for solitary pulmonary nodules. Zhong Nan Da Xue Xue Bao Yi Xue Ban 34:401–5, 2009
8. Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M: Bayesian networks for clinical decision support in lung cancer care. PloS one 8:e82349, 2013
9. Raja AS, Ip IK, Prevedello LM, Sodickson AD, Farkas C, Zane RD, et al: Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. Radiology 262:468–74, 2012
10. Canonical LLC. Ubuntu Linux. Available from: http://www.ubuntu.com. Accessed 25 December 2013.
11. DigitalOcean Inc. DigitalOcean. Available from: http://www.digitalocean.com. Accessed 25 December 2013.
12. NGINX Inc. NGINX. Available from: http://www.nginx.org. Accessed 25 December 2013.
13. Oracle Inc. MySQL. Available from: http://www.mysql.com. Accessed 25 December 2013.
14. Joyent Inc. NodeJS. Available from: http://www.nodejs.org. Accessed 25 December 2013.
15. Geisendorfer F. node-mysql. Available from: https://npmjs.org/package/mysql. Accessed 26 December 2013.
16. Otto M, Thornton J. Bootstrap. Available from: http://www.getbootstrap.com. Accessed 25 December 2013.
17. Google Inc. AngularJS. Available from: http://www.angularjs.org. Accessed 25 December 2013.
18. The JQuery Foundation. jQuery. Available from: http://www.jquery.com. Accessed 25 December 2013.
19. Kahn CE, Jr: Artificial intelligence in radiology: decision support systems. Radiographics 14:849–61, 1994
20. Boroczky L, Simpson M, Abe H, Drysdale J: Observer study of a prototype clinical decision support system for breast cancer diagnosis using dynamic contrast-enhanced MRI. AJR American J Roentgenol, 200:277–83, 2013
21. Wang KC, Jeanmenne A, Weber GM, Thawait SK, Carrino JA: An online evidence-based decision support system for distinguishing benign from malignant vertebral compression fractures by magnetic resonance imaging feature analysis. J Digit Imaging 24:507–15, 2011