An integrated clinical, CT, PFT database to better define an at risk population to screen for lung cancer

M. Khilkin, DO[1], C. Mattmann, PhD[2], P. Rivera, MD[1], J. Ko, MD[1], B. Phalan, BS[1] E. Eylers, RN[1], Sean Kelly, BS[2], D. Crichton, MS[2], and W.N. Rom, MD, MPH[1]. [1]New York University School of Medicine, NY. [2]Early Detection Research Network Project, Jet Propulsion Laboratory, CA. NIH U01CA86137, khilkm01@med.nyu.edu

BACKGROUND Routine CT screening for lung cancer is currently not recommended with studies only demonstrating a shift towards earlier-stage disease detected without a definitive mortality benefit. While CT screening is sensitive for identifying malignancy, its specificity can be poor, leading to unnecessary invasive procedures, irradiation from serial scans, patient anxiety and societal cost. One possible explanation for the failure of screening trials is very broad inclusion criteria, usually entirely based on a minimum age and smoking history. We developed a database integrating demographic, clinical, CT, PFT, environmental exposure information to better define subpopulations with risk factors for developing lung cancer.

METHODS The NYU Lung Cancer Biomarker Center has enrolled 1358 smokers and followed them prospectively for up to 6 years for the development of lung cancer. Participants answered a yearly supervised questionnaire that includes demographic information, ATS respiratory symptoms, occupational, family, medical, surgical and female-only history and underwent PFTs and MDCT scanning. Nodules were followed according to suspicion for lung cancer in either 3 and 12 or 6 and 12 months. Solid nodules were followed for at least 2 years and ground glass nodules were followed indefinitely yearly.

RESULTS In collaboration with NYU Radiology and EDRN/JPL we constructed an online MySQL database that allows retrospective and prospective follow-up of nodule size, 5 nodule and 6 adjacent parenchymal features, 121 thoracic and abdominal findings, 6 PFT and 700 questionnaire values. To construct the database, we are leveraging several emerging Java-based open source technologies, including: (1) Apache Tapestry for building a compact, easily maintainable web-based user interface to use for data input and for free-text and forms-based search, and (2) an object-relational persistence technology called Hibernate that affords us the ability to rapidly effect database design changes to our backend CT scan data model and collected data. Using these technologies, we are able to easily provide a means for persisting our collected data into MySQL, an open source relational database management system for further analysis.

CONCLUSIONS An online database allows the pooling of information to identify subpopulations more appropriate for CT screening for lung cancer.