Developing a New Low-Dose CT Image Library for Lung Cancer Screening

Paul Pinsky
Division of Cancer Prevention

NIH NATIONAL CANCER INSTITUTE

December 1, 2020

Problem

Low-dose Computed Tomography (LDCT) screening for lung cancer has been shown to reduce lung cancer mortality, however:

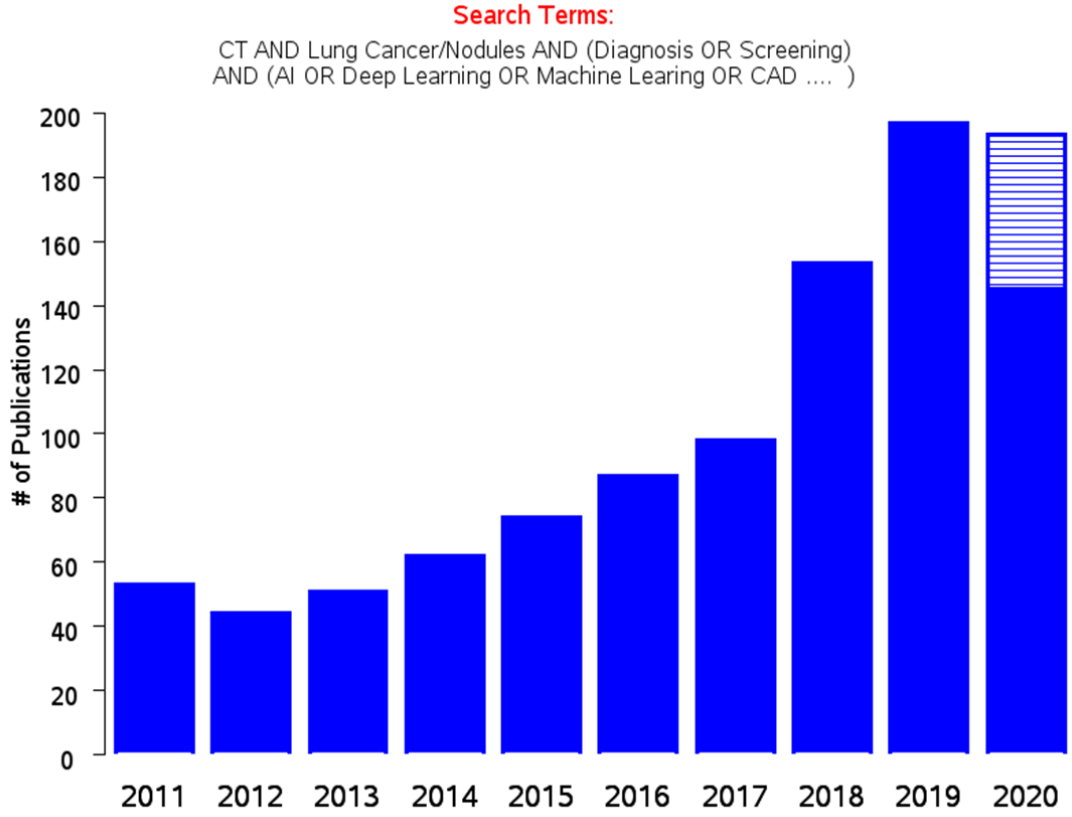
High false-positive rate (FPR) (~10-20%) results in anxiety, repeat imaging, and can lead to invasive diagnostic work-up

High FPR is a barrier to wider adoption of LDCT screening

Resolving the Problem Through Artificial Intelligence (AI)

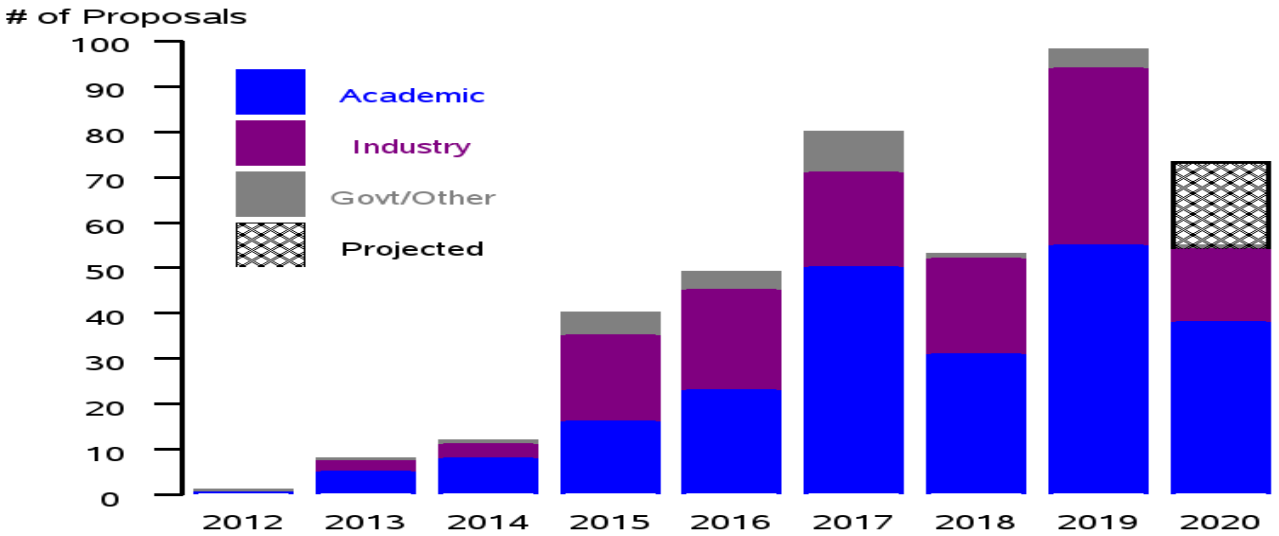
- To make screening more efficient and reduce screening-related harms, the FPR must be substantially reduced, while leaving test sensitivity essentially unchanged
- Approach to reducing FPR: development of artificial intelligence (AI) tools to assist radiologists in interpreting LDCT screening and diagnostic images
- AI development requires a large library of CT images and corresponding data

AI Research for Detection of Lung Nodules/Cancer



NLST CT Image Library

- The NLST CT image library is the main resource for the development of AI tools to predict lung cancer in LDCT screens
- Images available to the general research community
- 396 proposals for CT images (through 9/2020)



Shortcomings of the Existing NLST Image Library

- Images are from 2002-2006; CT technology has changed appreciably since then
- The NLST images have been widely disseminated, making true validation of AI algorithms problematic
- The NLST was a volunteer population and not representative of LDCT screening in current clinical practice
- 2020 USPSTF draft guidelines – Lower age, pack-years than NLST (age 50, 20 pack-years)
- The NLST image library has only screening LDCT exams, not diagnostic f/u CT exams

Request for Proposal (RFP)

- Create a new LDCT lung cancer screening image library
- Obtained with current LDCT technology and in standard clinical settings
- Also includes diagnostic f/u CT images
- Includes demographic, screening findings & clinical outcome (lung cancer incidence) data
- No enrolling/consenting of patients: only **retrospective** collection of de-identified images & data
- Make available to the research community
- Hold back a subset of images for algorithm validation (in coordination with FDA)

Addressing Prior BSA Concerns

- Concern 1: Assuring representativeness & generalizability of images
  - Images obtained from subjects undergoing routine screening, which is intended use population for AI algorithms
  - Facilities will be selected to insure geographic representativeness
  - Age distribution should reflect that of LDCT screening overall
  - Enrich for groups with high lung cancer risk (e.g., African Americans)

Addressing Prior BSA Concerns

- Concern 2: Assuring focus on reducing false positives
  - Collect adequate number of lung cancer associated images (~1500)
  - Enrich for false positive images – LungRADS 3-4 with no associated cancer (~9000)
  - Also include LungRADS 1-2 images (~4500)
  - Collect serial images to assess nodule change over time

Addressing Prior BSA Concerns

- Concern 3: Lack of Biospecimen Component
  - Image library is retrospective, so specimen collection not feasible
  - Resource with both CT exams and biospecimens useful
  - Leverage the existing NCI intramural Connect Cohort for this purpose

Proposed New Resource: CONNECT Cohort

CONNECT – New DCEG cohort

- Cancer etiology, early detection, risk prediction, survivorship
- 200,000 subjects aged adults 40-65 enrolled in integrative health care systems (IHCS)
- Serial questionnaires, serial biospecimen collection, EHR linkage

- Enroll sub-cohort of high-risk smokers to evaluate LDCT and biomarkers:
- Highly enriched for LDCT eligibility and past LDCT use
- Increase upper age limit (to 70)
- Coordinate blood draws with LDCT exams
- Use for other smoking-related cancers possible

CONNECT Sub-Cohort

- Collection of images & biospecimens for several hundred lung cancer cases over 5-7 years (and many more non-cases) potentially feasible
- Uses of sub-cohort:
  - For validated AI algorithms, assess the added value of incorporating biomarkers for screening/diagnostic work-up
  - Biomarker discovery
  - Correlation of biomarkers with CT features (e.g., ground glass)

www.cancer.gov

www.cancer.gov/espanol

Extra Slides

Timeline and Budget (Total Costs)

- Five-year contract
- Collection of data and images, set-up for image storage - up to 3 years
  [$ 4.5 million]
- Image storage and dissemination – up to 3 years
  [$ 0.5 million per year]
- Image validation set activities, including qualification as a Medical Device Development Tool (MDDT) with FDA – 2 years
  [~ $0.5-1 million]

- **Post-contract** – Plan to continue with image storage/dissemination in some manner

Scale of Project

	Proposed #
Unique Subjects *	15,000
Screening LDCT Images	22,500
Diagnostic CT Images	6,000-8000
Subjects with lung cancer-associated image (diagnosed within 18 months of a screen)	1500
Subjects with Lung-RADS positive screen (no cancer)	9000
Subjects with (only) Lung-RADS negative screens (no cancer)	4500

* Enrich for Racial/Ethnic Minorities, Other Special Populations