Breakout Session 1

Data Use in the Time of Screening and AI: Defining Consensus, Value, Context and Responsibilities

Quantitative Imaging Workup XVIII:

Optimizing Thoracic Imaging to Detect and Manage Early Lung Cancer/COPD

November 4-5, 2021 (Virtual)

Questions

1. What do we envision using big data for currently?
2. Assuming the infrastructure exists for supporting the data, curating, monitoring and releasing it, what are challenges is using it and putting it together
3. Mechanisms to continuously collect data versus single data dump.
4. Control on what data is used for and by whom
5. Measures of success
6. Who should do this

1. What do we envision using big data for currently?

- Focus of the database
  - lung cancer screening or
  - all the other possibilities, e.g. cardiac disease, emphysema etc.
- Make it as broad as possible
- Biggest value of lung cancer screening - > general health evaluation

- I-ELCAP as an example
  - Earlier papers focused on lung cancer screening-related topics
  - In recent years, starting to shift more towards other secondary findings

2. Assuming the infrastructure exists for supporting the data, curating, monitoring and releasing it, what are the challenges in using it and putting it together

- Existing resources:
  - I-ELCAP
  - MIDRC- seek out to institutions that have different populations to get representative groups
  - GO2 Foundation- over 800 centers Center of Excellence, around 50 continuum care centers.

- Challenges:
  - No uniform way to collecting data collection.
  - Each group/institution have different policies on how they want to share the data and what restrictions they want to put on the data
  - Need of critical information beyond CT images, e.g. clinical information such as medical history, physical & diagnostic workups, lab results, etc.
  - How do we reach out to all the screening sites out there?

2. Assuming the infrastructure exists for supporting the data, curating, monitoring and releasing it, what are the challenges in using it and putting it together (Continued)

- Population-based approach:
  - SEER-Medicare database
    - Links all fee-for-service data and additional links with research identifiable files: demographics, zipcode, addresses, prescriptions, diagnosis and procedure codes, and lab test results.
  - Link to commercial insurers
  - Allows for much richer longitudinal use
- Potential collaboration with the VA
  - Resources, technology, will and interest are all there
  - Standard health measures already stored within the VA corporate data warehouse
  - Availability of data beyond CT images – Million Veteran Program (MVP)

3. Mechanisms to continuously collect data versus single data dump

- What is the most important thing to get?
  - Images? Metadata? Annotated images?
- A database that continue to evolve
  - Keeping up with technology
  - Continue to update results (e.g. normal value)
- Challenges with curation on a continuous basis
  - Data cleaning

4. Control on what data is used for and by whom

- Who would have access?
  - Patient concerns
  - Institution concerns (may vary on whether industry can have access to it)
  - This can be a driver for start up companies in particular to help them accelerate in a meaningful way
- How could this be done?
  - Through NIH?
  - Outside NIH? Partner up with major organizations?
    - Prevent Cancer, GO2 Foundation
    - HHS-AHRQ
    - National Library of Medicine,
    - CMS ResDAC (U of Minnesota)
- Cost
  - Minor investment: Computer infrastructure
  - Major investment: Coordination (site oversight, contractual arrangements)

5. Measures of success

- To assess how beneficial it is, need to be able to track
  - who uses it,
  - #. of publications,
  - #. of products developed as a result of the use of data

6. Who should do this

- Who would support something like this?
  - Grants? Philanthropic donations?
  - Need to scope out the cost
  - Organizations to seek out:
    - Prevent Cancer foundation
    - GO2 Foundation
    - American Lung Association
    - QIBA