AI Data Readiness Challenge for the NCI Cancer Research Data Commons (CRDC)

Agnes McFarlin

October 17, 2024

NIH NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Project Design – Use Case and Resources

Use Case

- Category 2: Identify Cancerous lung nodules using the National Lung Screening Trial (NLST) data for early detection.

Data Commons Used

- Imaging Data Commons (IDC)

Study and Datasets Used

- The National Lung Screening Trial data set. (NLST)

Data Types

- Structured - (Patient Metadata)
- Unstructured - images (DICOM/JPEG)

Project Design - Methodology

NIH NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Methodology, continued

Pre-Processing Steps

- Select appropriate data.
- Add labels, shuffle data together.
- Change into a format more easily used for machine learning

Model Summary

- Model built using Pytorch in Python
- Accepts 512 X 512 Grayscale images as input
- Performs Binary Classification as output

Metrics Results

Metric	Value
Recall	0.58
F-1 Score	0.60
Precision	0.65

NIH NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Findings and Recommendations – Data Accessibility and Data Availability

Data Accessibility:

- There are several ways to download and work with data, but some ways are more advertised than others.
  - Some users may not be familiar with or be intimidated by having to use cloud resources such as Google Cloud products.

Recommendations:

- Improve data accessibility by increasing visibility of alternative tools such as the idc_index* package and make sure the package gets updated regularly.

Data Availability:

- All patient data relevant to my use case is publicly available, able to be previewed easily and does not require any special requests to be used.

Recommendations:

- None at this time.

https://doi.org/10.1148/rg.230180 *

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

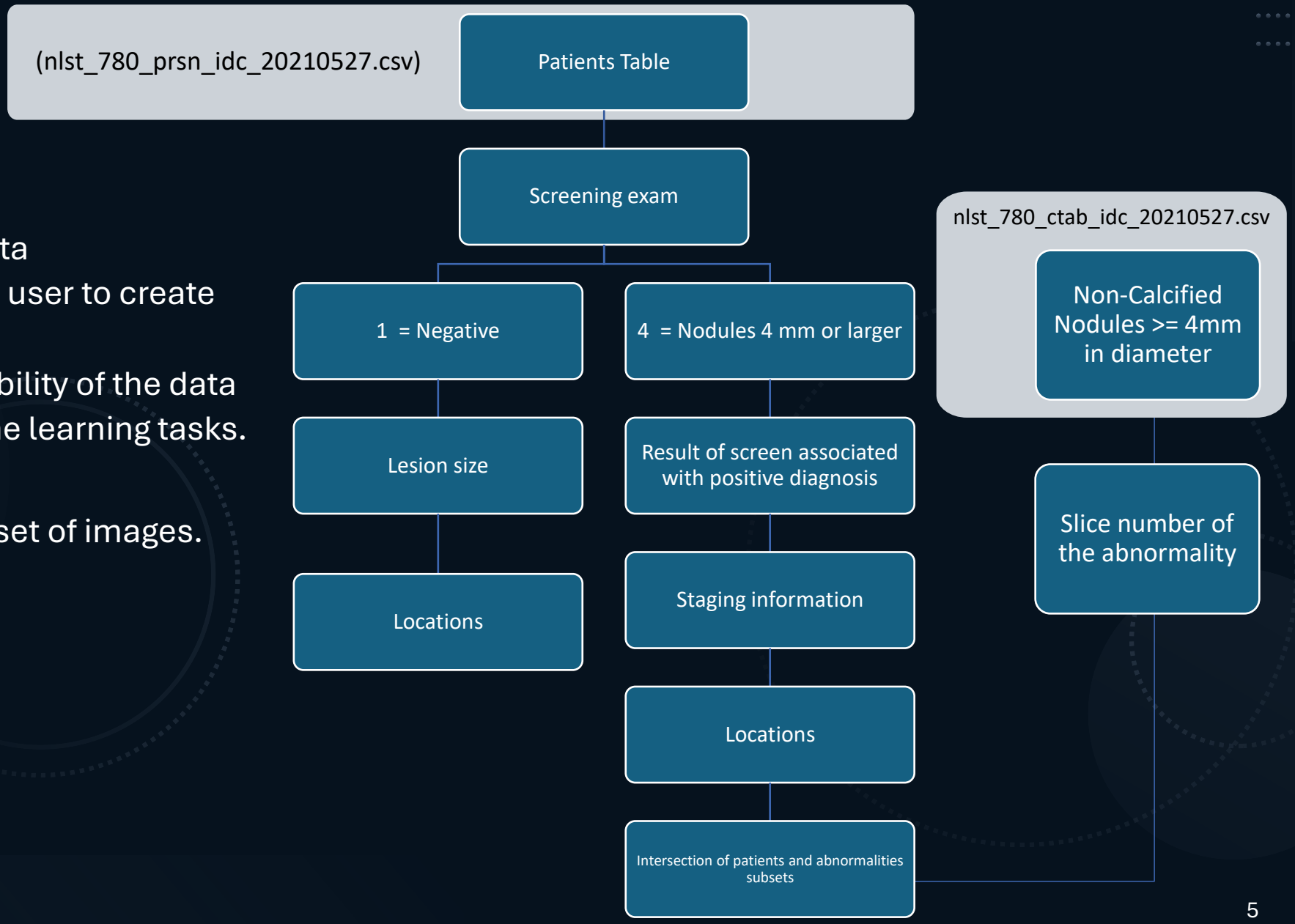
Findings and Recommendations - Accuracy

Accuracy:

- Lack of Annotated Data
  - Leaves it up to the user to create and define labels.
  - Limits the applicability of the data for certain machine learning tasks.

Recommendations:

- Create an annotated set of images.



Findings and Recommendations – Data Completeness and Data Privacy

Data Completeness:

- Of the patients that were queried there were no missing patient IDs. The IDs were able to be related to the metadata. Of the downloaded data there were also no duplicated patient IDs.

Recommendations

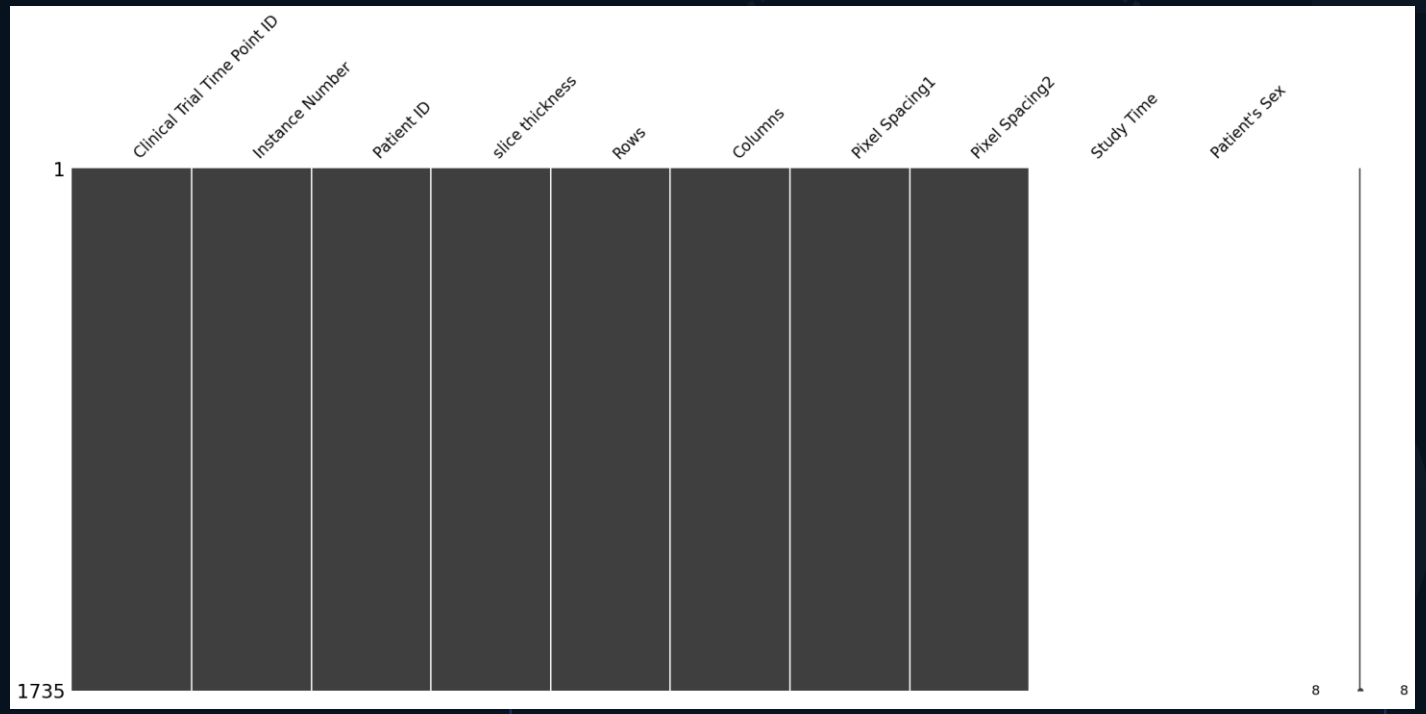
- None at this time.

Data Privacy:

- All patient records used were found to be properly anonymized, where applicable.
- Patient gender and study times were anonymized. Names and any identifying information were also anonymized.

Recommendations

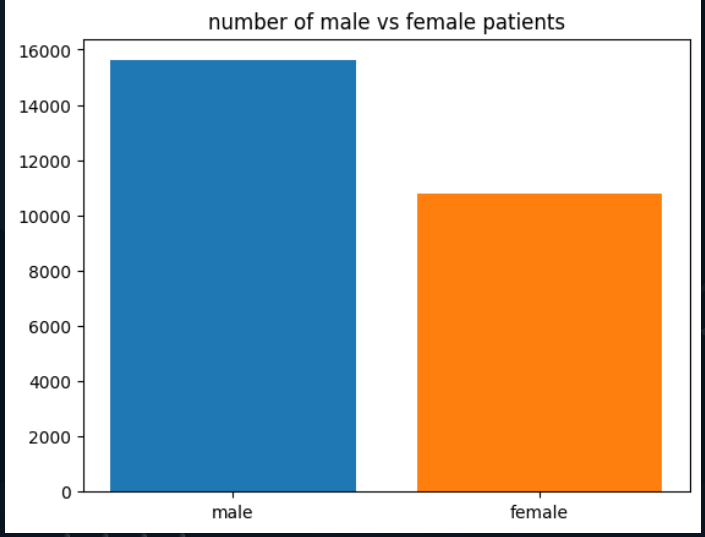
- None at this time.



Findings and Recommendations – Diversity

Diversity:

- The data set hosted by IDC could benefit from broader representation.
  - There were many more male participants than female. But there are still enough of each to create a balanced data set.
  - But most patients fall into one demographic category.
- Recommendations
  - Data enrichment might be possible by using other similar data sets.



Number of patients per demographic group for positive patient group
571
24
10
5
2

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Summary

The NLST is a great source of information but will become more AI ready if the issues mentioned are addressed.

- Improve data accessibility by making sure people are aware of all of the ways to download the data.
- Improve accuracy by adding annotations.
- Improve diversity by looking for other similar datasets containing needed data.

Thank you for your time