RSNA® Radiological Society of North America

ACR® American College of Radiology

AAPM American Association of Physicists in Medicine

MIDRC Medical Imaging and Data Resource Center.

Performance Metrics for Evaluation of COVID-19 AI: Technology Development Project (TDP) 3c

Michael McNitt-Gray, Ph.D.

Department of Radiological Sciences

David Geffen School of Medicine at UCLA

NIH National Institute of Biomedical Imaging and Bioengineering

David Geffen School of Medicine at UCLA

UCLA Health

The University of Chicago

Motivation

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

1) As machine intelligence (MI) algorithms are developed, appropriate means for benchmarking performance are needed.
    (a) So you are developing an algorithm related to some aspect of detecting or diagnosing COVID or assessing prognosis or response to treatment; how do you know how well the algorithm works?
    (b) Is the method you have developed better than other methods? How can you compare performance across algorithms (even your algorithm v1 to v2)?
2) Consistent performance metrics will be determined for specific MI tasks in order to evaluate new AI methods for improved health.

GOALS of MIDRC Technology Development Project (TDP) 3c

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

1) Develop useful information and recommendations on performance assessment metrics and analytical approaches that are task-specific; includes identifying analysis resources (research articles, available software packages, etc.).
2) Coordinate with TDP3d on how to best use sequestered data to provide reference analyses and benchmarks.
3) Coordinate with other MIDRC Collaborative Research Projects (CRPs) to ensure consistent use of performance assessment approaches/metrics.

Coordinate with CRPs

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

Project	Title
1	Natural Language Processing of Radiology Reports for COVID-19
2	Machine Intelligence Algorithms from Multi-Modal, Multi-institutional COVID-19 Data
3	Image Labeling and Annotation by a Crowd of Experts for COVID-19
4	Efficient Training and Explainability of Machine Learning Methods from Multi-Institutional Data
5	COVID Pneumonia Machine Learning Algorithm Validation and Visualization
6	Safe Public Training Dataset for COVID-19 Machine Learning Algorithms
7	Leveraging Registry Data to Conduct Virtual Clinical Trials
8	Prediction of COVID Pneumonia Outcome using Radiomic Feature Analysis
9	Radiomics & Machine Intelligence of COVID-19 for detection and diagnosis on chest radiographs and thoracic CTs
10	Visualization & Explainability of Machine Intelligence of COVID-19 for prognosis and monitoring therapy
11	Investigation of image-based biomarkers for radiogenomics of COVID-19
12	Determining COVID-19 image data quality, provenance, and harmonization

Recommendations Based on Clinical Task

- Example Tasks
  - Classification of patients as COVID-19 positive or negative based on CXR
    - Negative may include normals as well as other conditions such as pneumonias
  - Prediction of the severity of COVID-19 based on
    - CXR images
    - CT images
  - Automatic segmentation of COVID-19 areas of involvement based on CT
  - Predict short term prognosis based on CXR
    - Dismissal, hospital admission, ICU admission, intubation, etc.
  - Predict long term prognosis from other modalities/anatomic regions (brain MRIs, etc.)

- Recommendations will **NOT** cover all conceivable tasks or uses of the MIDRC database.
- Will evolve over time

Formulate Clinical Tasks as Analytical Tasks

- Example analytical tasks
  - Two-class classification
  - Multi-class classification
  - Estimation
  - Segmentation accuracy

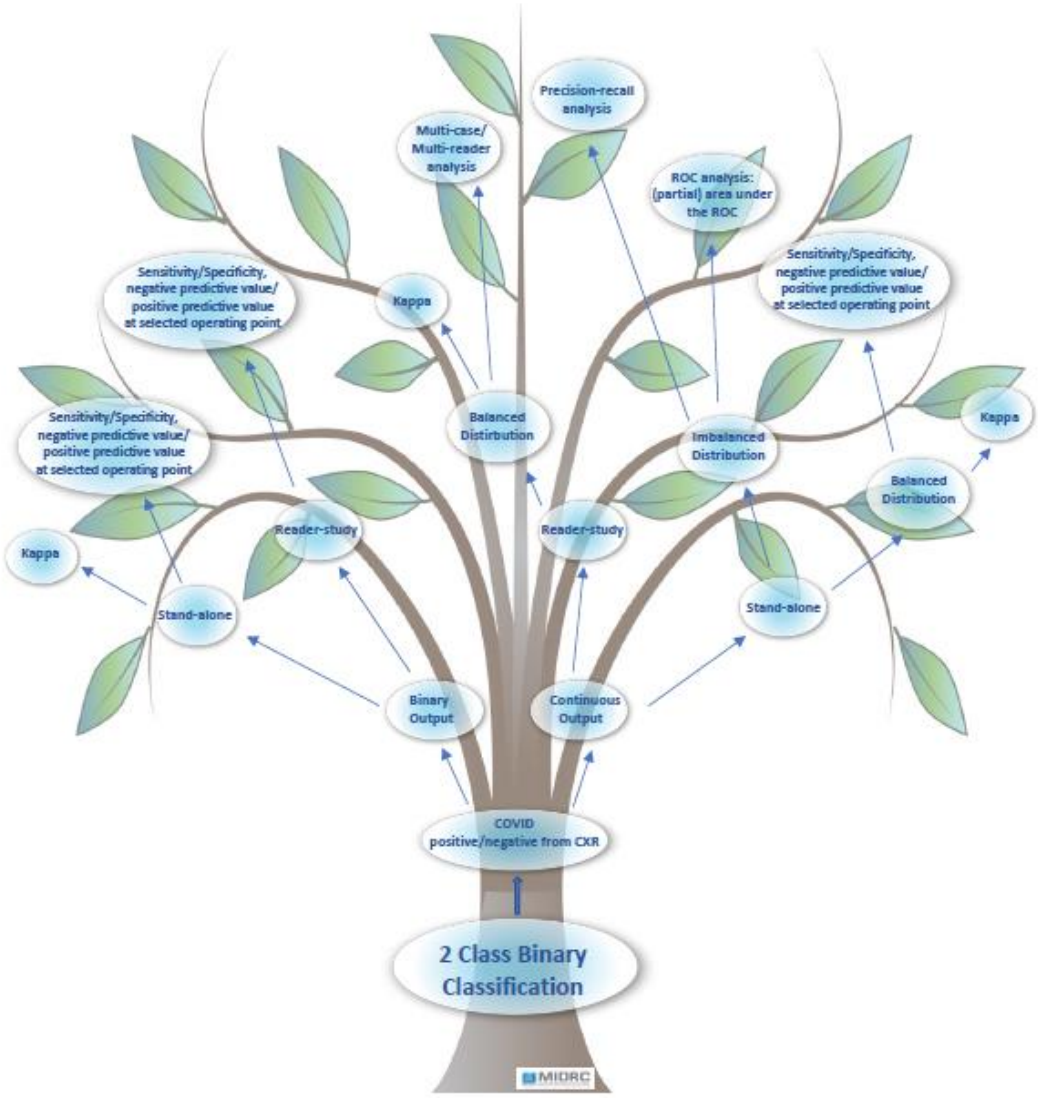
How to Get from Clinical Task to Analytical Task?

TDP 3c's First Goal:

Develop useful information and recommendations on performance assessment metrics and analytical approaches that are task-specific; includes identifying analysis resources (research articles, available software packages, etc.).

Our concept is to provide a decision tree to assist MIDRC users.

MIDRC TDP3c Decision Tree



MIDRC 3c Decision Tree

Provide the users with a series of questions

Based on responses, navigate tree

At end (“Leaf”), provide suggestions as to:

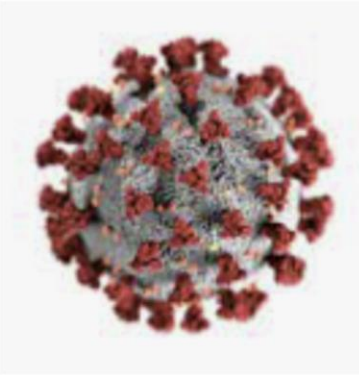
- Relevant literature
- Approach
- Metrics
- When available, software packages to carry out analysis

MIDRC TDP3c Decision Tree

Metrics, Technology Assessments and Reference Literature for various Clinical Tasks related to COVID-19

Welcome to the MIDRC metrics recommendation tree.

Give a summary of your task.



What type of task are you evaluating?

detection | diagnosis | prognosis | segmentation

MIDRC 3c Decision Tree

Metrics, Technology Assessments and Reference Literature for various Clinical Tasks related to COVID-19

Clinical task.

Please describe your clinical task.

Is your goal to distinguish between two clinical diagnoses?
(e.g. Distinguish between COVID + and COVID -)

Yes | No

MIDRC TDP3c Decision Tree

Metrics, Technology Assessments and Reference Literature for various Clinical Tasks related to COVID-19

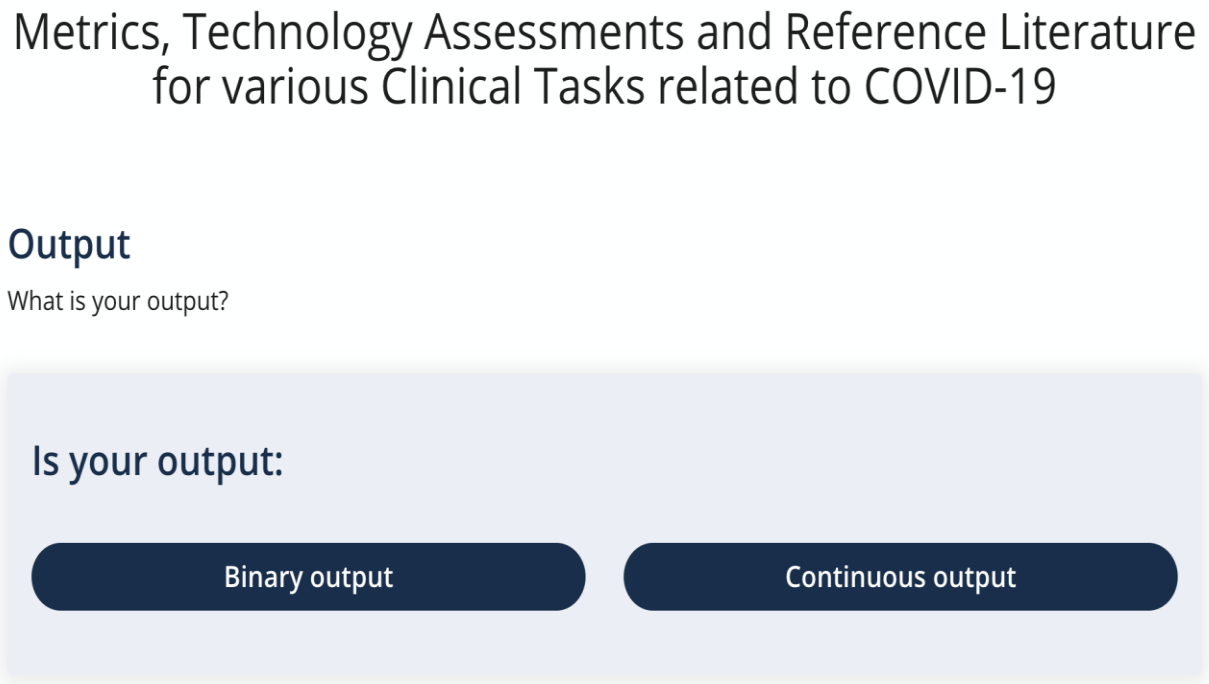
Modality

Describe your data.

What is the modality of your datatset?

Chest X-ray | Chest CT

MIDRC 3c Decision Tree



MIDRC TDP3c Decision Tree - Example

Q1. What type of task are you evaluating?

>> Diagnosis

Q2. Is your goal to distinguish between two clinical diagnoses? (e.g., "Determining COVID positive or negative")

>> Yes

Q3. What is the modality of your dataset?

>> CXR

Q4. Is your clinical output binary or continuous?

>> Continuous

Q5 might be: What is your reference/comparison standard?

>> PCR test result

MIDRC TDP3c Decision Tree - Example

Based upon the results of these questions, the decision tree may:

1. Identify the analytical task being performed (2-class classification task)
2. Provide some suggested analytical methods (e.g., ROC analysis)
3. Provide suggested performance metrics (e.g., area under ROC curve - AUC)
4. Provide links to software tools (online repositories, Githubs) to assist in performing the suggested analyses
   - Software analysis tools will provide a point estimate of the suggested performance metric as well as a confidence interval based on the sources of variability.

GOALS of MIDRC TDP3c

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

1) Develop useful information and recommendations on performance assessment metrics and analytical approaches that are task-specific; includes identifying analysis resources (research articles, available software packages, etc.).
2) Coordinate with TDP3d on how to best use sequestered data to provide reference analyses and benchmarks.
3) Coordinate with CRPs to ensure consistent use of performance assessment approaches and metrics.

Summary

MIDRC will facilitate AI device innovation, validation and

- Is available to developers of COVID-19 AI algorithms around the world
- Provides large, diverse data sets for training and testing
- Will develop tools to help users identify approaches to evaluate performance
- Allows for fair and robust comparisons between algorithms/approaches
  - Across a range of clinical tasks (detection, diagnosis, estimation, etc.)
- Will promote scientific “Challenges” and effective use of MIDRC data – both public and sequestered

Natalie Baughan, UChicago
Karen Drukker, UChicago
Maryellen Giger, UChicago
Mike McNitt Gray, UCLA
Kyle Myers, FDA
Berkman Sahiner, FDA
Emily Townley, AAPM
Heather Whitney, UChicago/Wheaton

MIDRC Performance Metric and Benchmarking Team (TDP3c)

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.