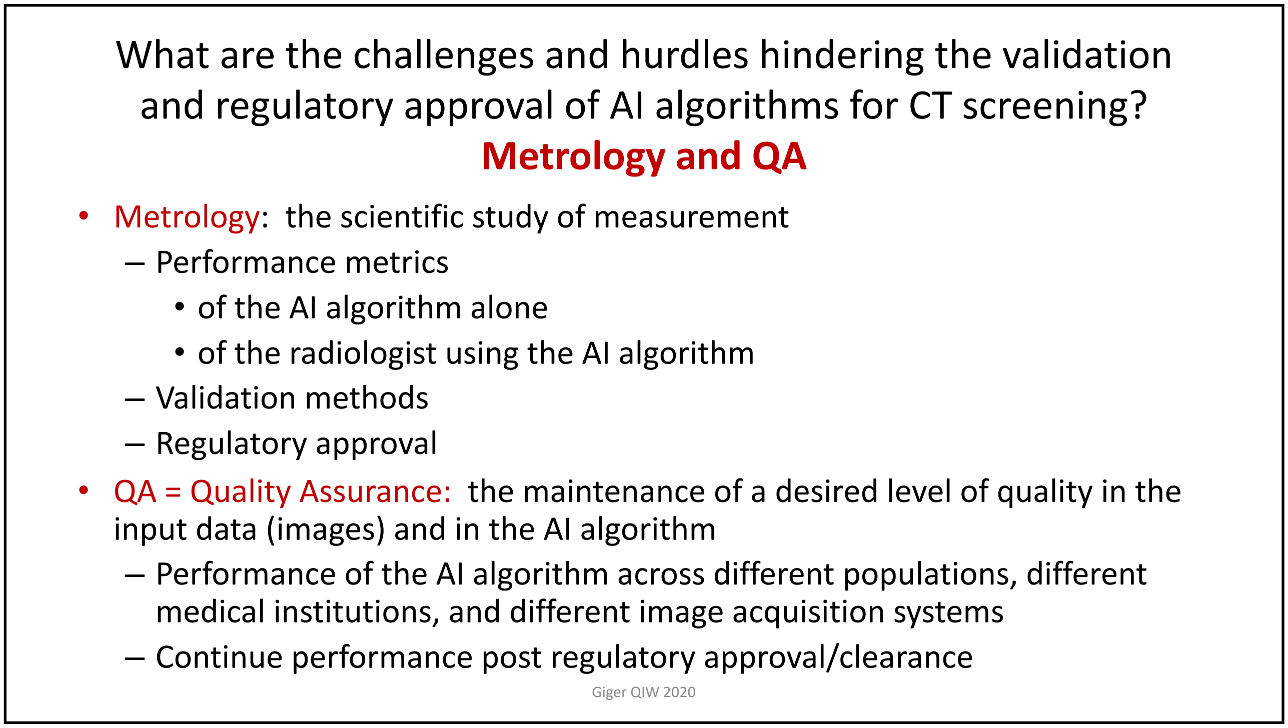
Prevent Cancer Foundation
Quantitative Imaging Workshop XVII

Metrology and QA Perspective: What are the challenges and hurdles hindering the validation and regulatory approval of AI algorithms for CT screening?

Maryellen L. Giger, PhD
A. N. Pritzker Professor of Radiology / Medical Physics
The University of Chicago
m-giger@uchicago.edu

Funding and COIs

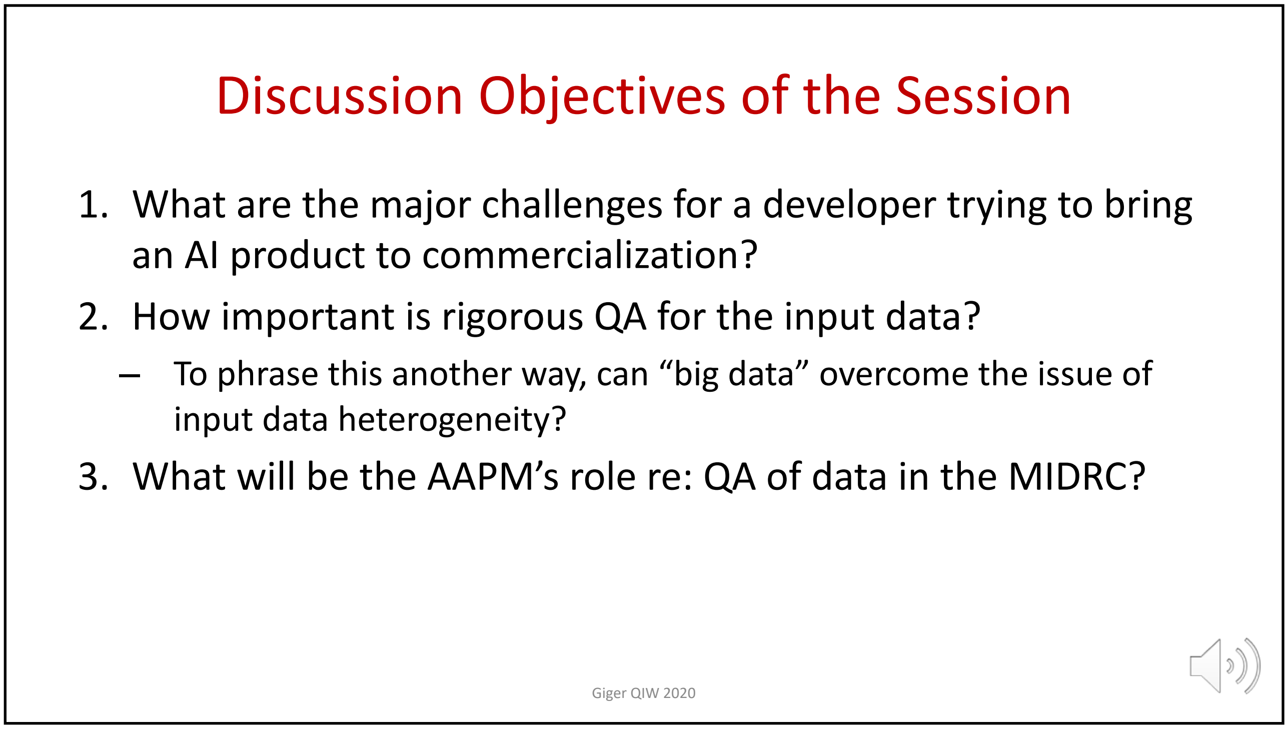
- Supported in parts by NIH grants CA 195564, CA 166945, and CA 189240; NIH S10 OD025081 Shared Instrument Grant; and CTSA UL1 TR000430; UChicago Cancer Center Koleseiki Funding and Dancing with Chicago Celebrities Funding; CDAC Grant; c3.ai Grant; NIBIB COVID-19 Contract 75N92020D00021
- MLG is a stockholder in R2/Hologic, shareholder in Qview, and receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi, and Toshiba.
- MLG is scientific advisor, co-founder, and equity holder in Quantitative Insights, [now Qlarity Imaging] makers of QuantX -- the first FDA-cleared machine learning system for aiding in cancer diagnosis.
- It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.



What are the challenges and hurdles hindering the validation and regulatory approval of AI algorithms for CT screening?

Metrology and QA

- Metrology: the scientific study of measurement
  - Performance metrics
    - of the AI algorithm alone
    - of the radiologist using the AI algorithm
  - Validation methods
  - Regulatory approval
- QA = Quality Assurance: the maintenance of a desired level of quality in the input data (images) and in the AI algorithm
  - Performance of the AI algorithm across different populations, different medical institutions, and different image acquisition systems
  - Continue performance post regulatory approval/clearance



Discussion Objectives of the Session

1. What are the major challenges for a developer trying to bring an AI product to commercialization?
2. How important is rigorous QA for the input data?
   - To phrase this another way, can "big data" overcome the issue of input data heterogeneity?
3. What will be the AAPM's role re: QA of data in the MIDRC?

Medical Image Interpretation

Medical images are meaningless grayscale/colorscale patterns unless "viewed and analyzed" by an intelligent observer

- Radiologist, Computer (AI), or Combination of human & computer (AI-aided)

Tasks of the Human eye-brain system

- Finding/locating a signal in an image
- Characterizing/classifying/diagnosing the signal as disease or non-disease
- Clinical decision making on patient management through integrated diagnostics (monitoring)

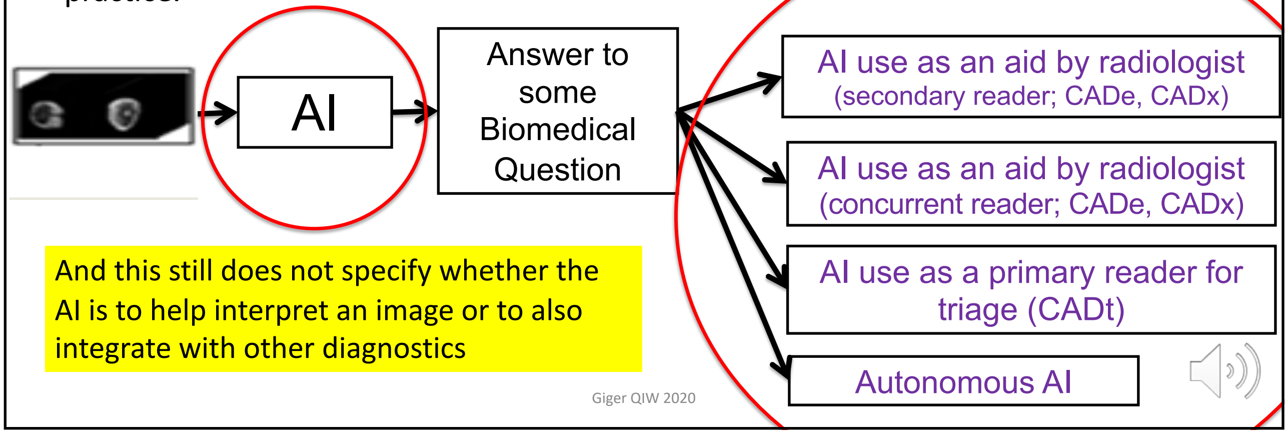
Tasks of AI (computer vision, radiomics, machine learning, deep learning)

- Similar – converting images to quantitative values
- Need to know the clinical task!!!!

Giger QIW 2020

In assessing challenges for translation and commercialization, need to know the clinical task and how the AI will be used

- **AI is the computer algorithm** (e.g., radiomics/machine learning/DL) that exists in a system for various types of implementations in clinical practice – CAD or otherwise.
- CAD (computer-aided diagnosis) describes a method of **how the AI is used** in clinical practice.



AI

Answer to some Biomedical Question

AI use as an aid by radiologist (secondary reader; CADe, CADx)

AI use as an aid by radiologist (concurrent reader; CADe, CADx)

AI use as a primary reader for triage (CADt)

Autonomous AI

And this still does not specify whether the AI is to help interpret an image or to also integrate with other diagnostics

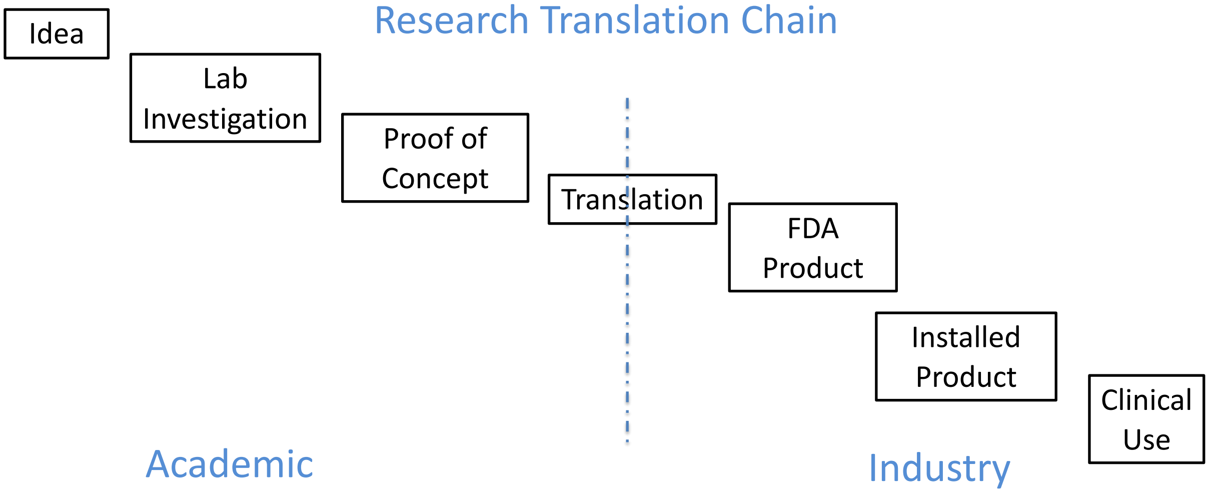
Giger QIW 2020

Discussion Objectives of the Session

1. What are the major challenges for a developer trying to bring an AI product to commercialization?
2. How important is rigorous QA for the input data? (To phrase this another way, can "big data" overcome the issue of input data heterogeneity?)
3. What will be the AAPM's role re: QA of data in the MIDRC?

Scientific research creates new knowledge, however, translation to the public often requires collaborations of academia & industry

Research Translation Chain



Idea

Lab Investigation

Proof of Concept

Translation

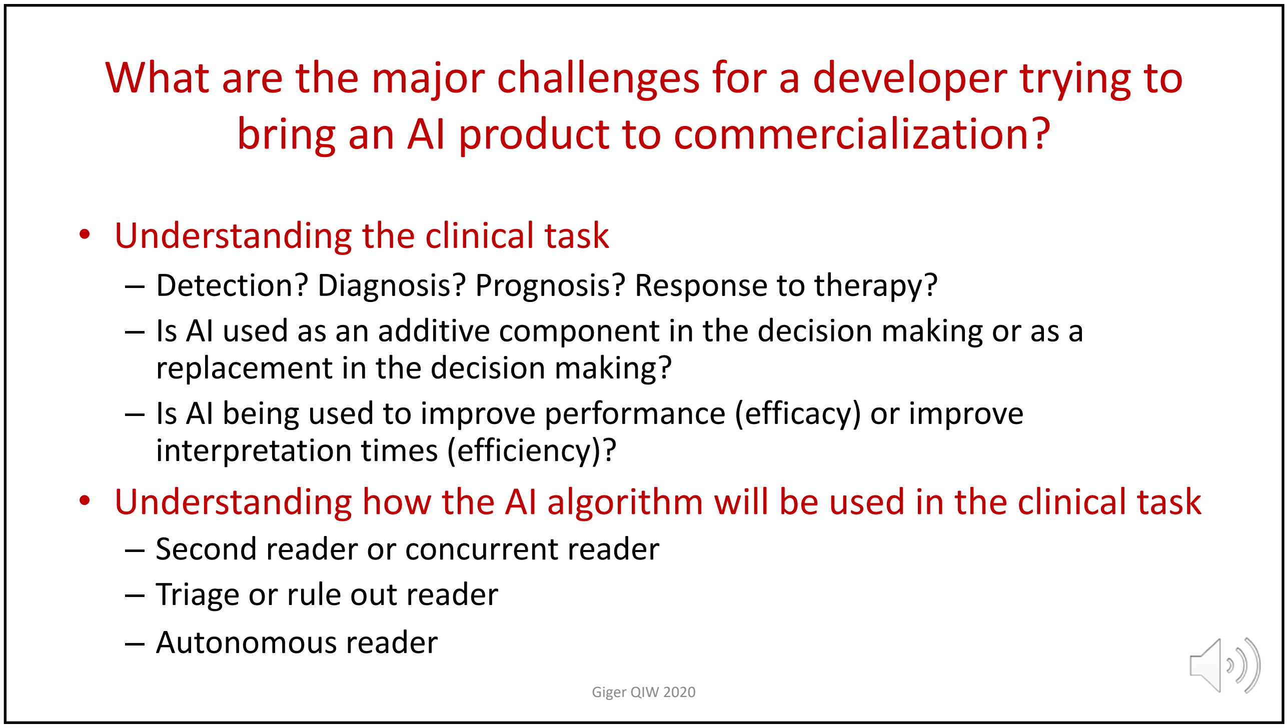
FDA Product

Installed Product

Clinical Use

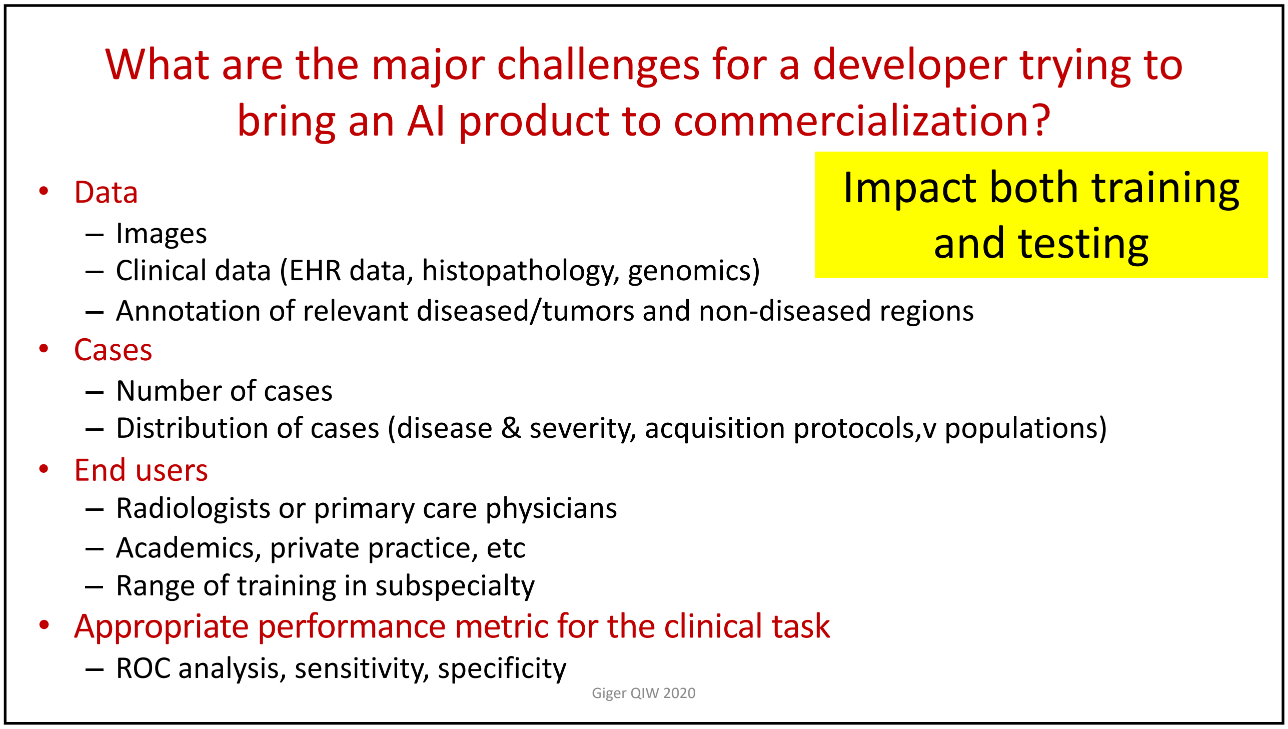
Academic

Industry



What are the major challenges for a developer trying to bring an AI product to commercialization?

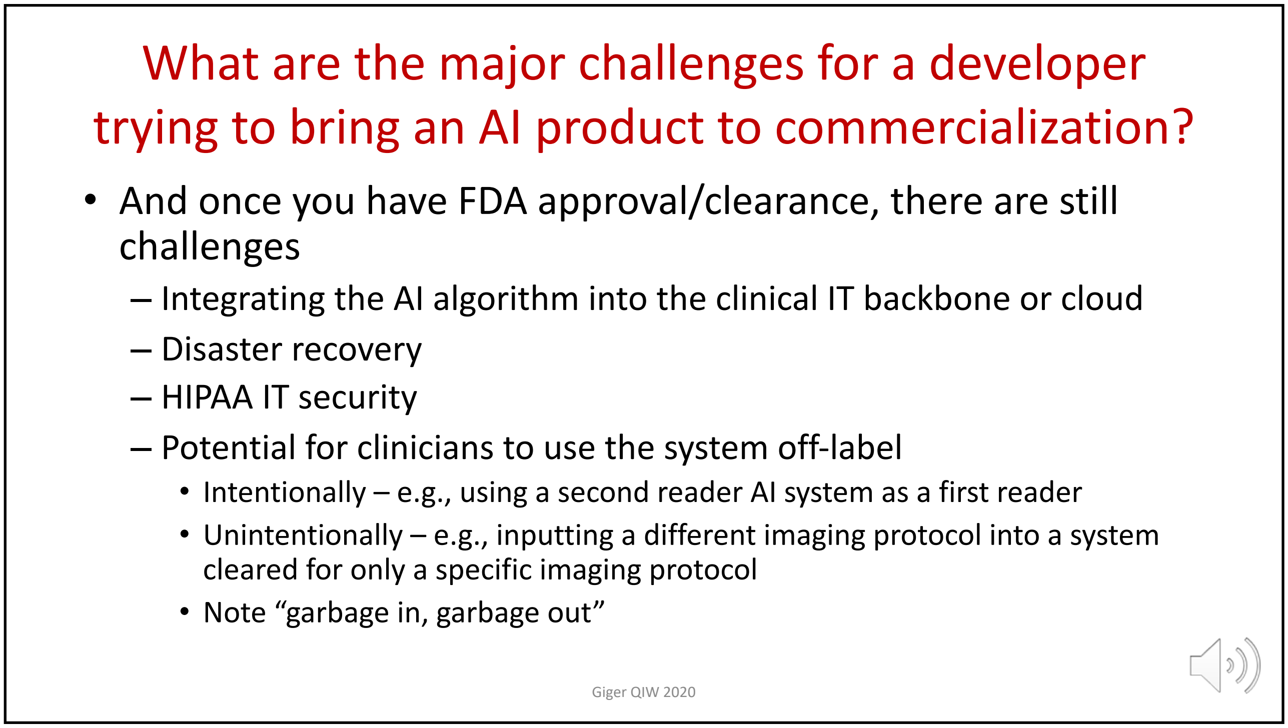
- Understanding the clinical task
  - Detection? Diagnosis? Prognosis? Response to therapy?
  - Is AI used as an additive component in the decision making or as a replacement in the decision making?
  - Is AI being used to improve performance (efficacy) or improve interpretation times (efficiency)?
- Understanding how the AI algorithm will be used in the clinical task
  - Second reader or concurrent reader
  - Triage or rule out reader
  - Autonomous reader



What are the major challenges for a developer trying to bring an AI product to commercialization?

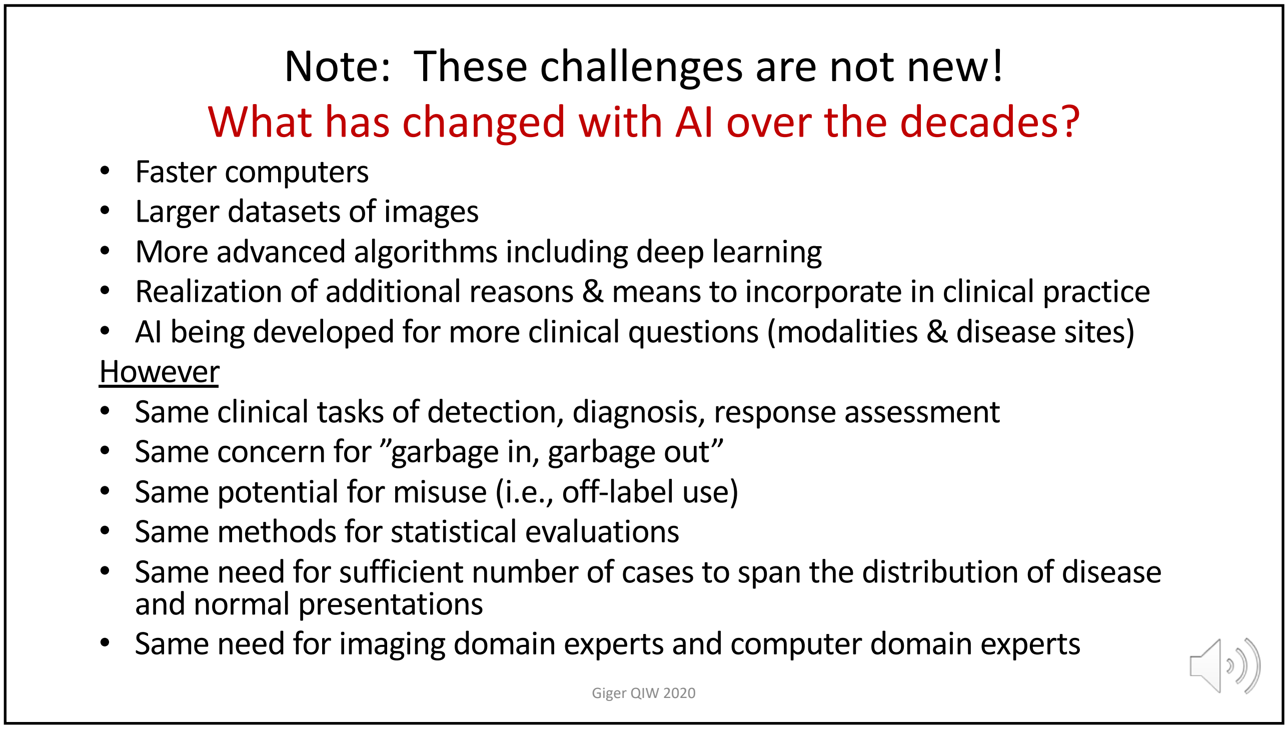
Impact both training and testing

- Data
  - Images
  - Clinical data (EHR data, histopathology, genomics)
  - Annotation of relevant diseased/tumors and non-diseased regions
- Cases
  - Number of cases
  - Distribution of cases (disease & severity, acquisition protocols,v populations)
- End users
  - Radiologists or primary care physicians
  - Academics, private practice, etc
  - Range of training in subspecialty
- Appropriate performance metric for the clinical task
  - ROC analysis, sensitivity, specificity



What are the major challenges for a developer trying to bring an AI product to commercialization?

- And once you have FDA approval/clearance, there are still challenges
  - Integrating the AI algorithm into the clinical IT backbone or cloud
  - Disaster recovery
  - HIPAA IT security
  - Potential for clinicians to use the system off-label
    - Intentionally – e.g., using a second reader AI system as a first reader
    - Unintentionally – e.g., inputting a different imaging protocol into a system cleared for only a specific imaging protocol
    - Note “garbage in, garbage out”



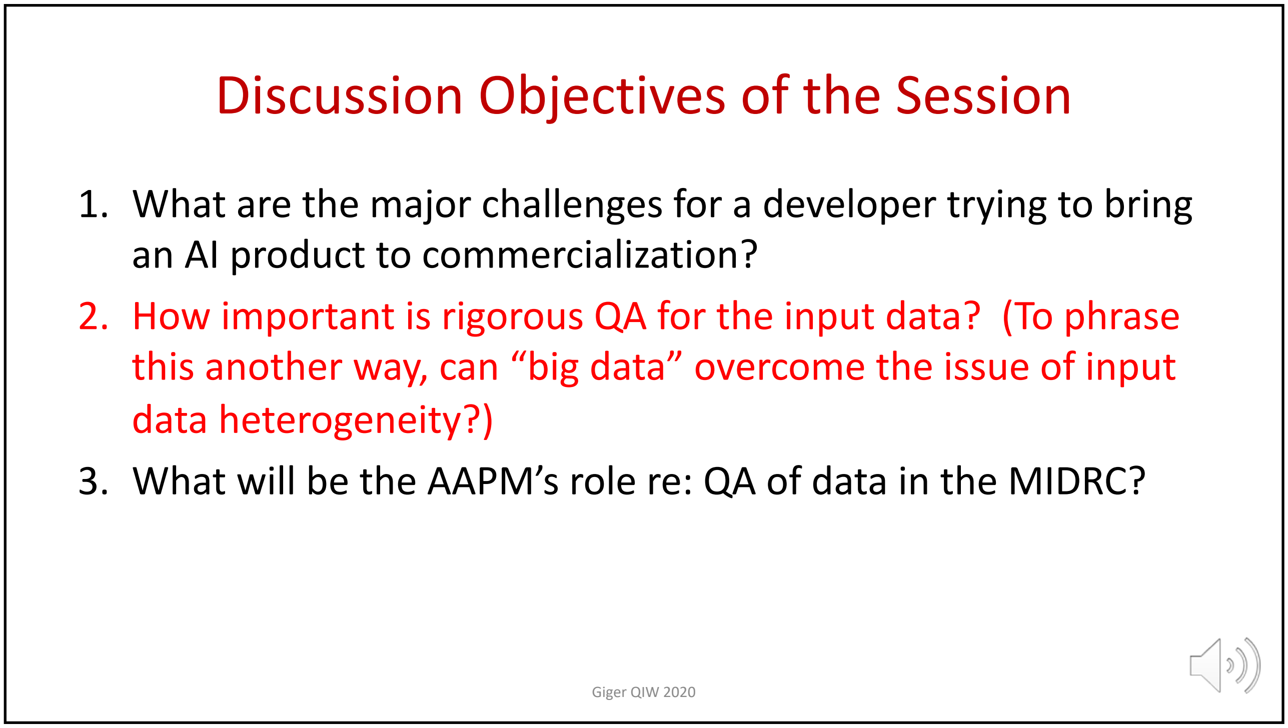
Note: These challenges are not new!

What has changed with AI over the decades?

- Faster computers
- Larger datasets of images
- More advanced algorithms including deep learning
- Realization of additional reasons & means to incorporate in clinical practice
- AI being developed for more clinical questions (modalities & disease sites)

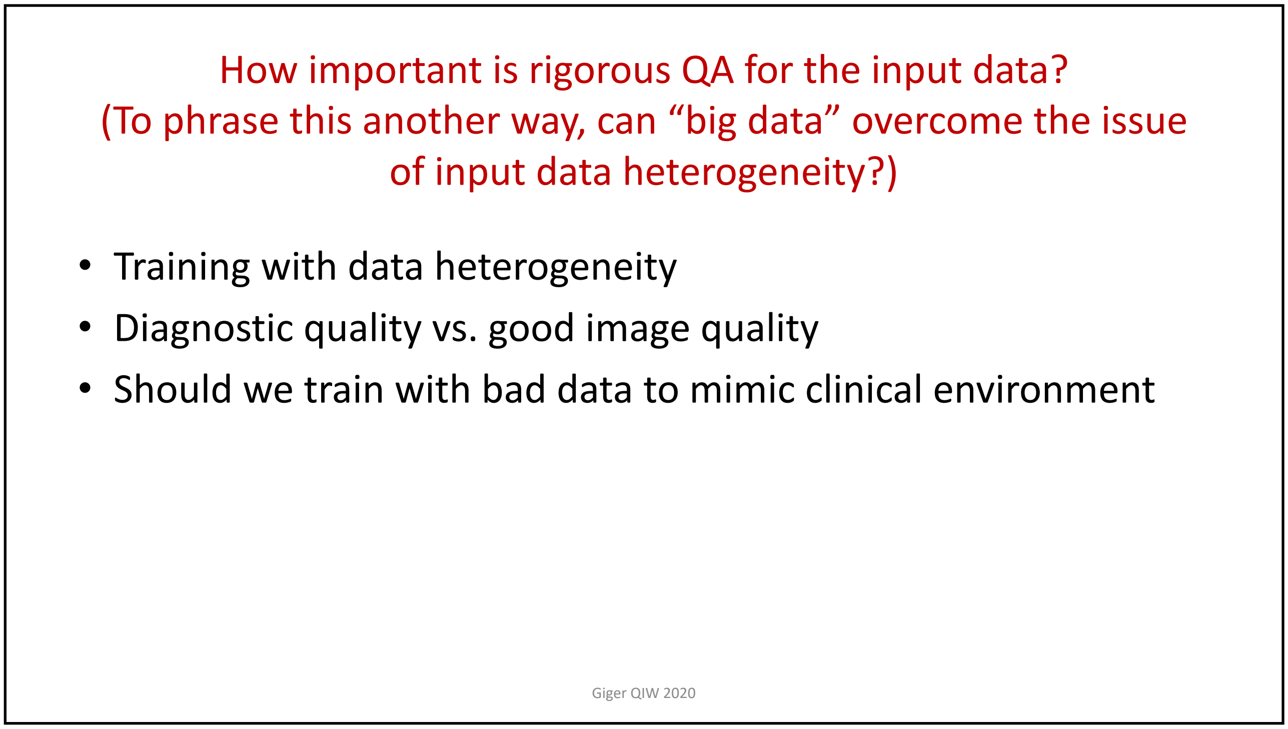
<u>However</u>

- Same clinical tasks of detection, diagnosis, response assessment
- Same concern for ”garbage in, garbage out”
- Same potential for misuse (i.e., off-label use)
- Same methods for statistical evaluations
- Same need for sufficient number of cases to span the distribution of disease and normal presentations
- Same need for imaging domain experts and computer domain experts



Discussion Objectives of the Session

1. What are the major challenges for a developer trying to bring an AI product to commercialization?
2. How important is rigorous QA for the input data? (To phrase this another way, can "big data" overcome the issue of input data heterogeneity?)
3. What will be the AAPM's role re: QA of data in the MIDRC?



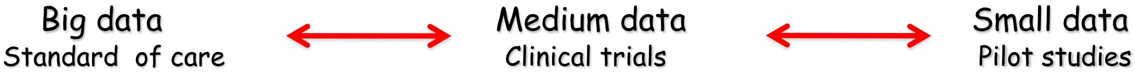
How important is rigorous QA for the input data? (To phrase this another way, can "big data" overcome the issue of input data heterogeneity?)

- Training with data heterogeneity
- Diagnostic quality vs. good image quality
- Should we train with bad data to mimic clinical environment

Robustness: Databases & QA

Differences in image acquisition or the population may affect computer-extracted image-based phenotypes

- Manufacturer
- Imaging protocol
- Geographic location
  - Racial differences in disease prevalence and characteristics
- Actual outcome data such as survival may not be available and intermediate alternatives may need to be used


Big data
Standard of care
Medium data
Clinical trials
Small data
Pilot studies

If the sample does not accurately represent the population of interest, statistics are not meaningful.

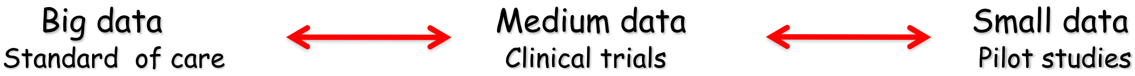
Giger QIW 2020

Robustness: Databases & QA

Differences in **image acquisition** or the population may affect computer-extracted image-based phenotypes

- Spatial resolution
- Noise
- System gain

- Manufacturer
- Imaging protocol
- Geographic location
  - Racial differences in disease prevalence and characteristics
- Actual outcome data such as survival may not be available and intermediate alternatives may need to be used


Big data
Standard of care
Medium data
Clinical trials
Small data
Pilot studies

If the sample does not accurately represent the population of interest, statistics are not meaningful.

Giger QIW 2020

Robustness: Databases & QA

Differences in image acquisition or the **population** may affect computer-extracted image-based phenotypes

- Manufacturer
- Imaging protocol
- Geographic location
  - Racial differences in disease prevalence and characteristics
- Actual outcome data such as survival may not be available and intermediate alternatives may need to be used

- Disease prevalence
- Geographical
- Single vs. multiple institutions
- Lack of harmonization

Big data ↔ Medium data ↔ Small data
Standard of care ↔ Clinical trials ↔ Pilot studies

If the sample does not accurately represent the population of interest, statistics are not meaningful.

Robustness: Databases & QA

Differences in image acquisition or the population may affect computer-extracted image-based phenotypes

- Manufacturer
- Imaging protocol
- Geographic location
  - Racial differences in disease prevalence and characteristics
- Actual outcome data such as survival may not be available and intermediate alternatives may need to be used

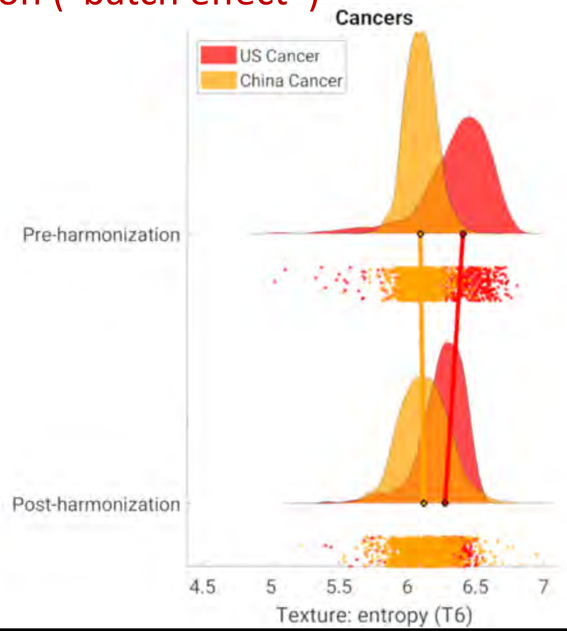
- Size and distribution of dataset will depend on task
- Quality of annotations; truth

Big data ↔ Medium data ↔ Small data
Standard of care ↔ Clinical trials ↔ Pilot studies

If the sample does not accurately represent the population of interest, statistics are not meaningful.

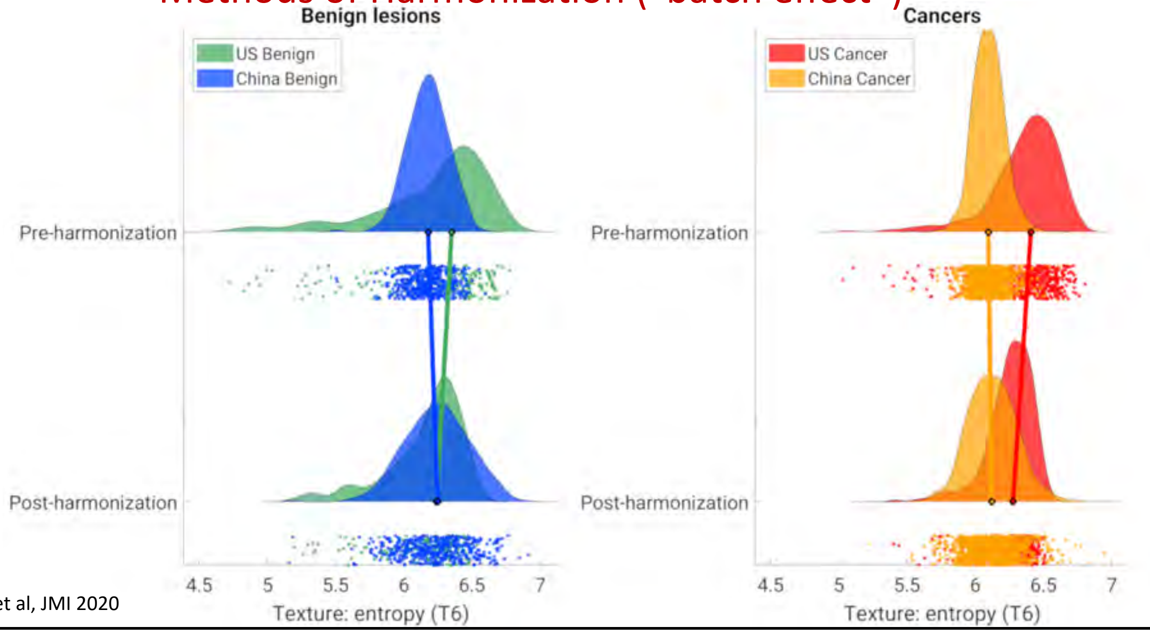
How can “Big Data” Compensate for Heterogeneities? Methods of Harmonization (“batch effect”)

- Harmonized breast MRI radiomic features from a US dataset and a China dataset
- ComBat data harmonization used to standardize radiomic features across populations according to the feature categories dependence upon system gain, resolution, and noise (Johnson et al., *Biostatistics.* 2007)
- Uses empirical Bayes methods to pool information across lesions in each population to shrink batch effect parameter estimates of mean and variance



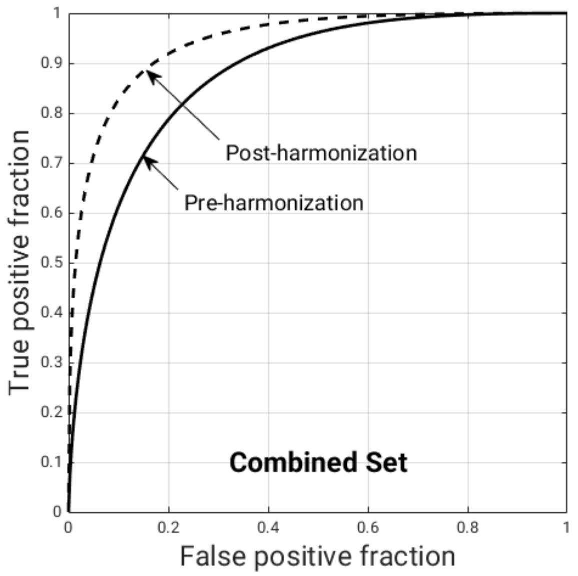
Whitney H, Li H, Ji Y, Liu P, Giger ML: Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. J. Med. Imag. 7(1), 012707, doi: 10.1117/1.JMI.7.1.012707, 2020

How can “Big Data” Compensate for Heterogeneities? Methods of Harmonization (“batch effect”)

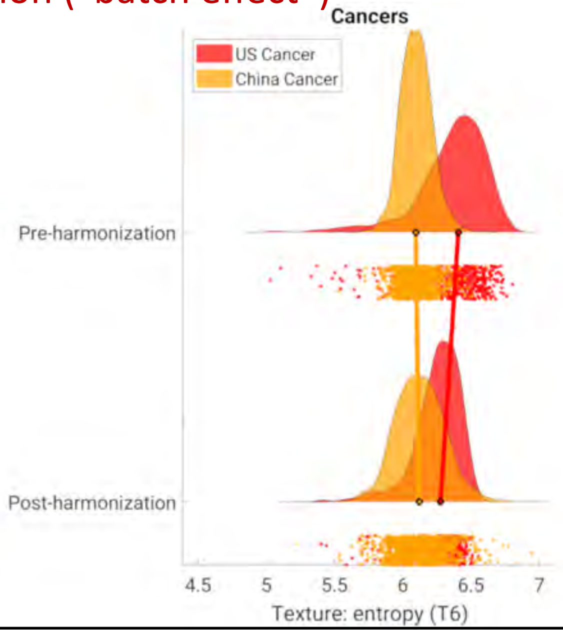


Whitney et al, JMI 2020

How can “Big Data” Compensate for Heterogeneities? Methods of Harmonization (“batch effect”)



Whitney et al. JMI 2020



How can “Big Data” Compensate for Heterogeneities? Methods of Harmonization (“batch effect”)

- Think like a human
- If harmonization does not work, might need AI “tuned” to specific populations, acquisition systems, etc.
- Recall discussions on QIBA

Discussion Objectives of the Session

1. What are the major challenges for a developer trying to bring an AI product to commercialization?
2. How important is rigorous QA for the input data? (To phrase this another way, can “big data” overcome the issue of input data heterogeneity?)
3. What will be the AAPM’s role re: QA of data in the MIDRC?

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

MIDRC.org

NIH National Institute of Biomedical Imaging and Bioengineering

THE UNIVERSITY OF CHICAGO

GEN3 DATA COMMONS

AMERICAN ASSOCIATION of PHYSICISTS IN MEDICINE

RSNA Radiological Society of North America

ACR AMERICAN COLLEGE OF RADIOLOGY

Overview: The COVID-19 pandemic presents an urgent and critical public health crisis. Essential biomedical research and development is needed to urgently address:

(i) **surveillance and early detection** of COVID-19 resurgence via monitoring of imaging and other clinical data

(ii) **detection, triaging, and differential diagnosis** of COVID-19 patients

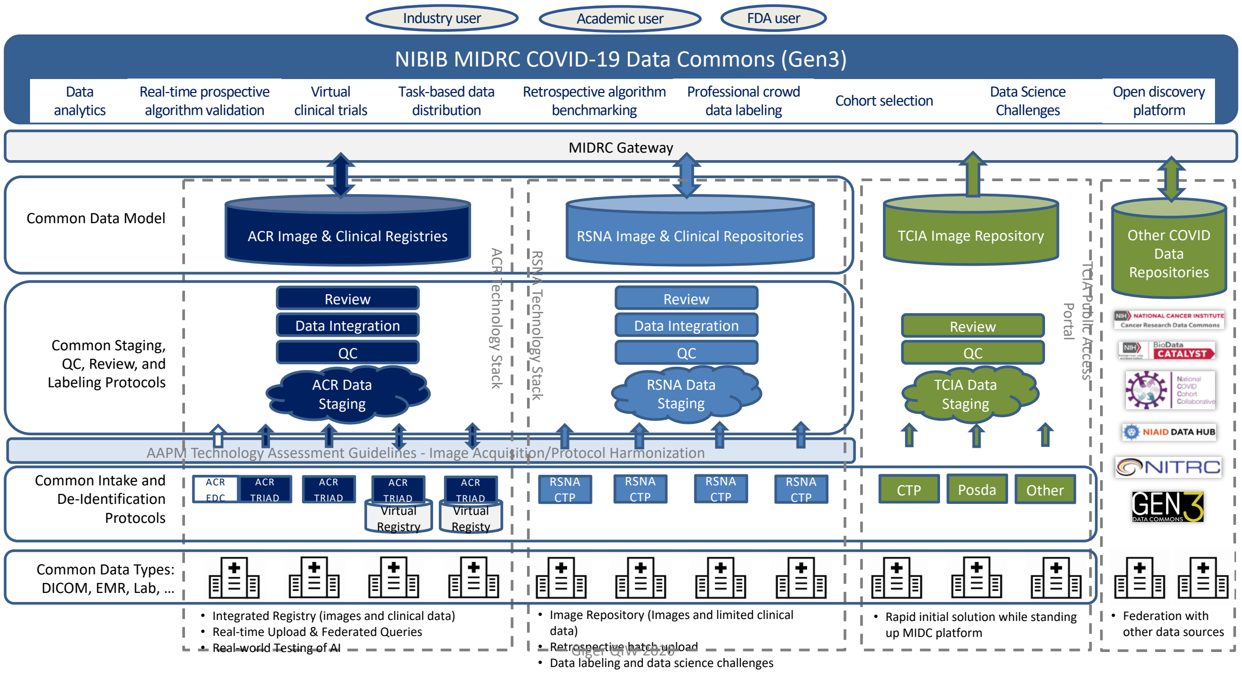
(iii) **prognosis, including prediction and monitoring of response,** for use in patient management.

In response to this need, representatives of the RSNA, ACR, and AAPM with NIBIB have jointly developed the Medical Imaging and Data Resource Center (MIDRC) for rapid and flexible **collection, AI research, and dissemination** of imaging and associated data, to be administered and hosted through the University of Chicago.

Two Major Scientific Components of MIDRC

A. Open Discovery Data Repository
creation, testing, quality assurance, and data connectivity

B. Machine Intelligence Computational Capabilities
clinically relevant algorithms and software tools



Industry user
Academic user
FDA user

NIBIB MIDRC COVID-19 Data Commons (Gen3)

Data analytics
Real-time prospective algorithm validation
Virtual clinical trials
Task-based data distribution
Retrospective algorithm benchmarking
Professional crowd data labeling
Cohort selection
Data Science Challenges
Open discovery platform

MIDRC Gateway

Common Data Model
ACR Image & Clinical Registries
RSNA Image & Clinical Repositories
TCIA Image Repository
Other COVID Data Repositories

Common Staging, QC, Review, and Labeling Protocols
Review
Data Integration
QC
ACR Data Staging
ACR Technology Stack
RSNA Technology Stack
Review
Data Integration
QC
RSNA Data Staging
Review
QC
TCIA Data Staging
TCIA Public Access Portal

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons
NIH BioData CATALYST
National COVID Cohort Collaborative
NIAID DATA HUB
NITRC
GEN3 DATA COMMONS

AAPM Technology Assessment Guidelines - Image Acquisition/Protocol Harmonization

Common Intake and De-Identification Protocols
ACR EDC
ACR TRIAD
ACR TRIAD
ACR TRIAD Virtual Registry
ACR TRIAD Virtual Registy
RSNA CTP
RSNA CTP
RSNA CTP
RSNA CTP
CTP
Posda
Other

Common Data Types: DICOM, EMR, Lab, …

- Integrated Registry (images and clinical data)
- Real-time Upload & Federated Queries
- Real-world Testing of AI

- Image Repository (Images and limited clinical data)
- Retrospective batch upload
- Data labeling and data science challenges

- Rapid initial solution while standing up MIDC platform

- Federation with other data sources

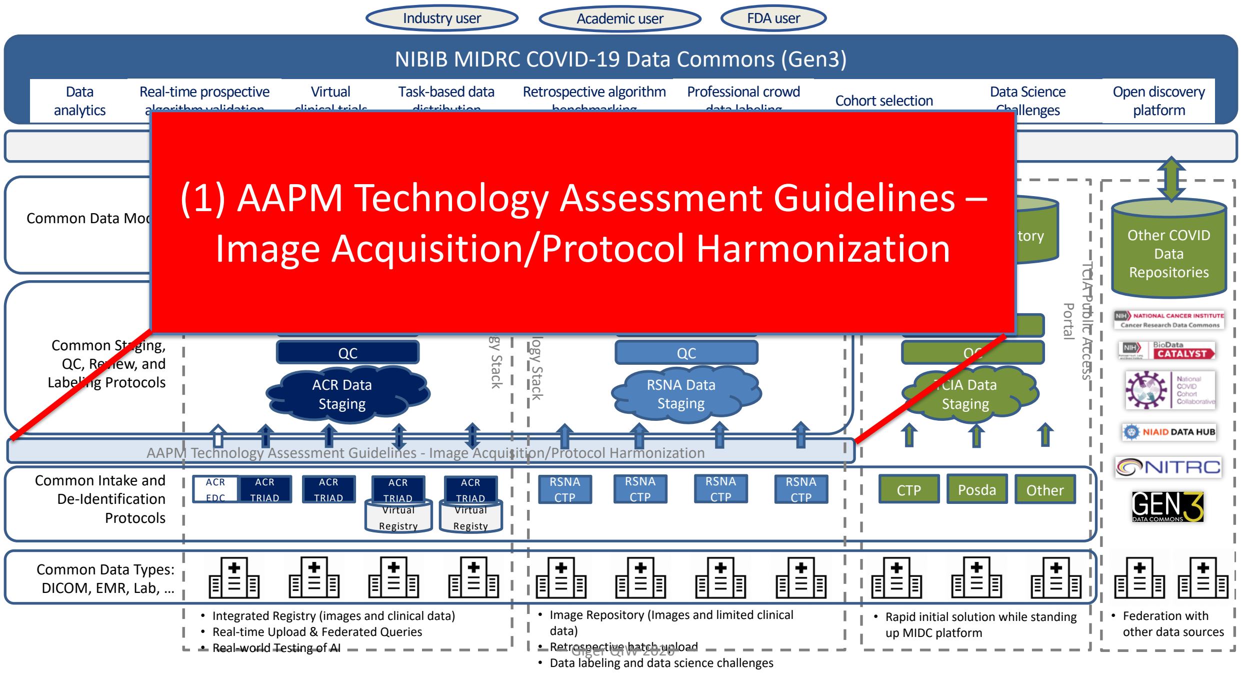
Giger QIW 2020

MIDRC: Technology Development Projects

The **MIDRC infrastructure and processes** will be created through five **Technology Development Projects**, which will be conducted collaboratively:

1. Creating an open discovery platform for COVID-19 imaging and associated data (led by RSNA).
2. Creating a real-world testing and implementation platform with direct real-time connections to health care delivery organizations (led by ACR).
3. Developing and implementing quality assurance and evaluation procedures for usage across the MIDRC (led by AAPM).
4. Enabling data intake, access and distribution via a world-facing data commons portal (led by all three plus Gen3).
5. Linking the MIDRC to other clinical and research data registries (led by all three plus Gen3).

Giger QIW 2020



Industry user

Academic user

FDA user

NIBIB MIDRC COVID-19 Data Commons (Gen3)

Data analytics | Real-time prospective algorithm validation | Virtual clinical trials | Task-based data distribution | Retrospective algorithm benchmarking | Professional crowd data labeling | Cohort selection | Data Science Challenges | Open discovery platform

(1) AAPM Technology Assessment Guidelines – Image Acquisition/Protocol Harmonization

Common Data Mod

Common Staging, QC, Review, and Labeling Protocols

QC

ACR Data Staging

QC

RSNA Data Staging

QC

TCIA Data Staging

tory

gy Stack

logy Stack

TCIA Public Access Portal

Other COVID Data Repositories

NATIONAL CANCER INSTITUTE Cancer Research Data Commons

BioData CATALYST

National COVID Cohort Collaborative

NIAID DATA HUB

NITRC

GEN3 DATA COMMONS

AAPM Technology Assessment Guidelines - Image Acquisition/Protocol Harmonization

Common Intake and De-Identification Protocols

ACR EDC | ACR TRIAD | ACR TRIAD | ACR TRIAD Virtual Registry | ACR TRIAD Virtual Registy

RSNA CTP | RSNA CTP | RSNA CTP | RSNA CTP

CTP | Posda | Other

Common Data Types: DICOM, EMR, Lab, …

- Integrated Registry (images and clinical data)
- Real-time Upload & Federated Queries
- Real-world Testing of AI

- Image Repository (Images and limited clinical data)
- Retrospective batch upload
- Data labeling and data science challenges

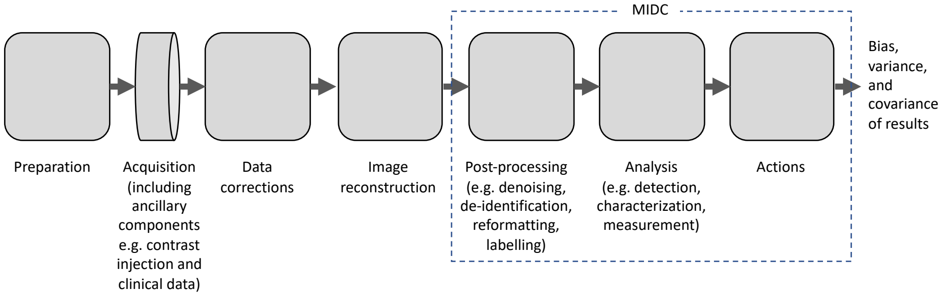
- Rapid initial solution while standing up MIDC platform

- Federation with other data sources

Giger QIW 2020

Developing and implementing quality assurance and evaluation procedures for usage across the MIDRC (led by AAPM)

1. Development of digital and physical imaging phantoms for COVID data
2. Assessment of image quality on ingestion into MIDRC



MIDC

Preparation

Acquisition (including ancillary components e.g. contrast injection and clinical data)

Data corrections

Image reconstruction

Post-processing (e.g. denoising, de-identification, reformatting, labelling)

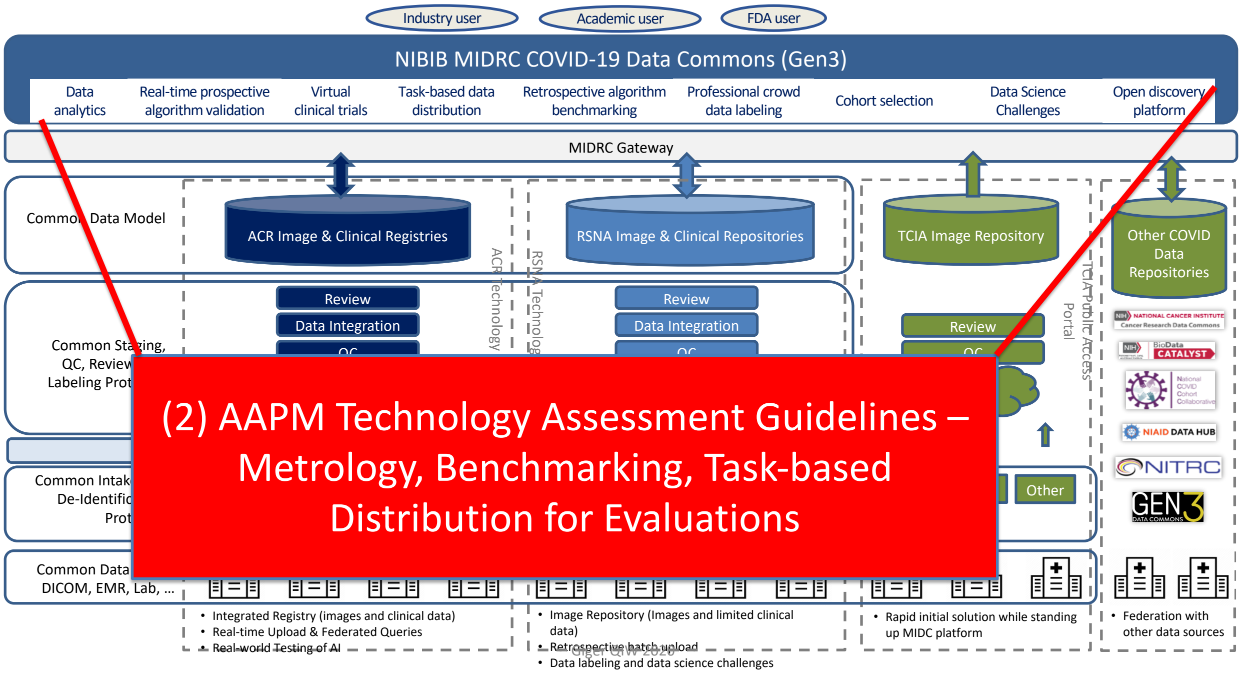
Analysis (e.g. detection, characterization, measurement)

Actions

Bias, variance, and covariance of results

P Kinahan, J Boone

Giger QIW 2020


Industry user
Academic user
FDA user
NIBIB MIDRC COVID-19 Data Commons (Gen3)
Data analytics
Real-time prospective algorithm validation
Virtual clinical trials
Task-based data distribution
Retrospective algorithm benchmarking
Professional crowd data labeling
Cohort selection
Data Science Challenges
Open discovery platform
MIDRC Gateway
Commo Data Model
ACR Image & Clinical Registries
RSNA Image & Clinical Repositories
TCIA Image Repository
Other COVID Data Repositories
Common Staging, QC, Review Labeling Prot
Review
Data Integration
QC
ACR Technology
RSNA Technolog
TCIA Public Access Portal
NATIONAL CANCER INSTITUTE Cancer Research Data Commons
BioData CATALYST
National COVID Cohort Collaborative
NIAID DATA HUB
NITRC
GEN3 DATA COMMONS
Common Intak De-Identific Prot
Other
Common Data DICOM, EMR, Lab, …
Integrated Registry (images and clinical data)
Real-time Upload & Federated Queries
Real-world Testing of AI
Image Repository (Images and limited clinical data)
Retrospective batch upload
Data labeling and data science challenges
Rapid initial solution while standing up MIDC platform
Federation with other data sources
Giger QIW 2020

(2) AAPM Technology Assessment Guidelines – Metrology, Benchmarking, Task-based Distribution for Evaluations

Developing and implementing quality assurance and evaluation procedures for usage across the MIDRC (led by AAPM)

- Development of benchmarking methods for the various technology assessment and clinical tasks in COVID-19 research and translation
- Development of task-based distribution methods
  - To facilitate FDA clearance and accelerate clinical usage, methods of selecting clinical cases for the independent testing are required to appropriately evaluate a validated algorithm.
  - For a specific clinical task, (e.g., diagnosis of COVID-19 from other presentations of pneumonia), a distribution of cases, matched to the clinical claim and intended patient population, will be randomly drawn from the larger sequestered dataset within the MIDRC, and subsequently used in the testing.

M Giger, M McNitt-Gray, B Sahiner, K Drukker, K Myers

Giger QIW 2020

Discussion Objectives of the Session

1. What are the major challenges for a developer trying to bring an AI product to commercialization?
2. How important is rigorous QA for the input data?
   - To phrase this another way, can “big data” overcome the issue of input data heterogeneity?
3. What will be the AAPM’s role re: QA of data in the MIDRC?

In summary, the way to use AI in clinical interpretation has been expanding beyond serving as a “second reader” and this directly affects how to evaluate the AI system (metrology & QA)

- Evaluation of the AI algorithm alone in the **particular task**
- Evaluation of the **end-user** using the output of the AI algorithm in the particular task
- Evaluation of the robustness over range of data quality
- Having the clinically-proven algorithm **used as intended**
- High-performing **AI used for the wrong clinical task** or incorrectly by the user will not yield expected clinical outcomes
- Need to realize that the **AI output is not always correct**.
  - What level is acceptable for clinical use?
  - Will vary with the tasks – cost/benefit
  - What level is high enough so explainability is not needed?

Thank you to Giger Lab

Recent & Current Graduate Students

Joel Wilkie, PhD
Martin King, PhD
Nick Gruszauskas, PhD
Yading Yuan, PhD
Robert Tomek, MS
Neha Bhooshan, PhD
Andrew Jamieson, PhD
Hsien-Chi Kuo, PhD
Martin Andrews, PhD
William Weiss, PhD
Chris Haddad, PhD
Natasha Antropova, PhD
Adam Sibley, PhD
Kayla Robinson, PhD
Jennie Crosby, PhD
Isabelle (Qiyuan) Hu
Jordan Fuhrman
Lindsay Douglas
Natalie Baughan

Research Lab

Karen Drukker, PhD
Hui Li, PhD
Heather Whitney, PhD
Yu Ji, MD
Chun Wai Chan, MS
Li Lan, MS
John Papaioannou, MS
Sasha (Alexandra) Edwards, MA
Madeleine Durkee, PhD
Summer medical students,
undergraduates, and
high school students

Collaborators

Gillian Newstead, MD
Hiro Abe, MD
Deepa Sheth, MD
Marcus Clark, MD
Yuan Ji, PhD
Greg Karczmar, PhD
Milica Medved, PhD
Yulei Jiang, PhD