

# Centralized Imaging Collaborations for AI Readiness

**Paul Kinahan, PhD**

**Vice-Chair of Radiology, Research**

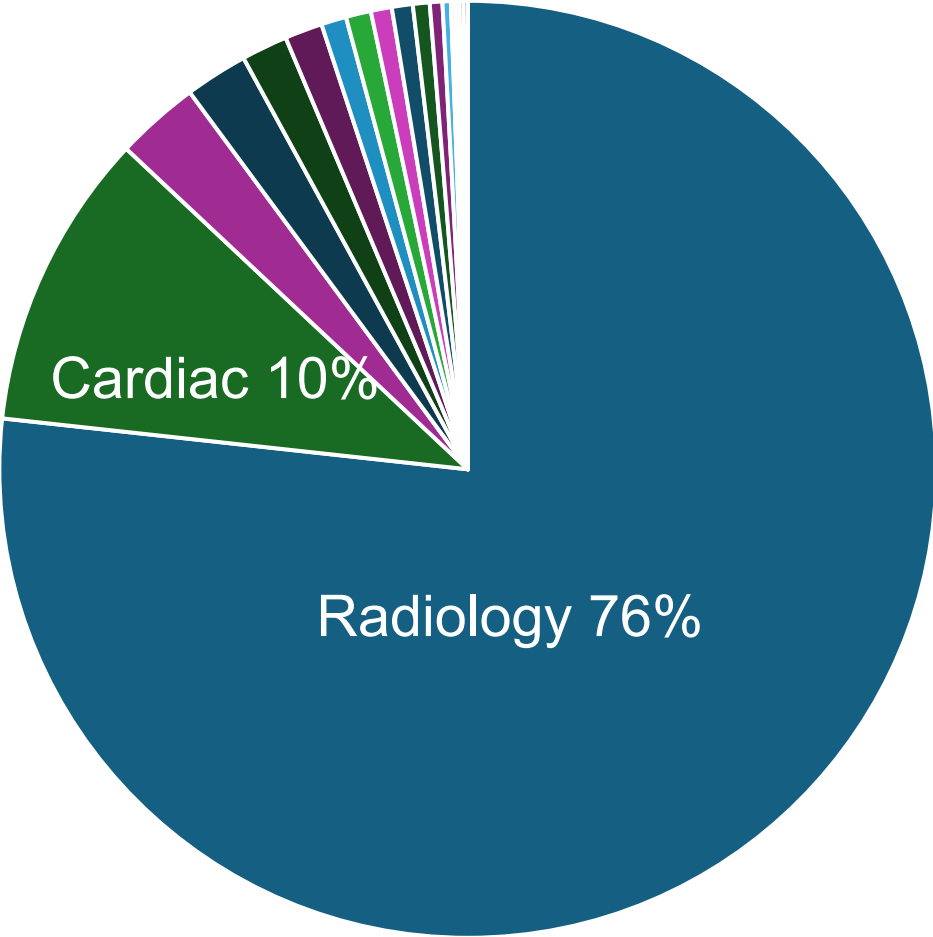
**University of Washington**

# Acknowledgments

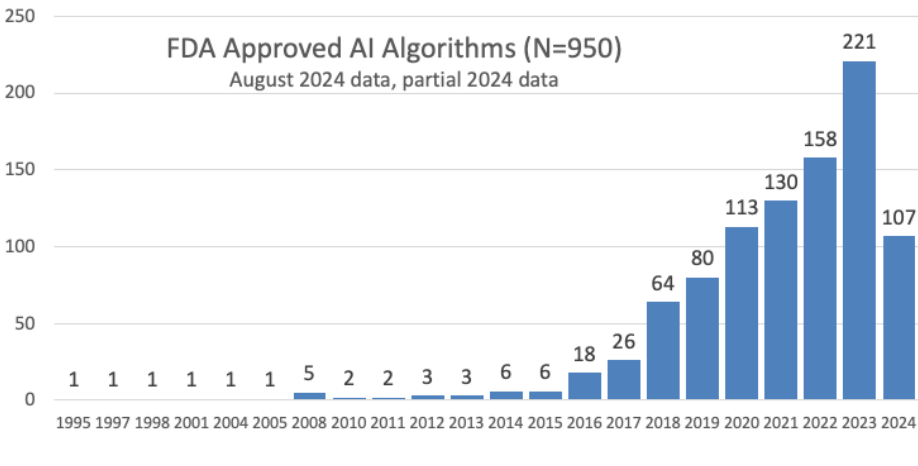
- NIH-NIBIB Contract 75N92020D00021
- The MIDRC leadership team
- Members of the MIDRC Data Quality and Harmonization Subcommittee
- Members of Technology Development Project 3: Develop and implement quality assurance and evaluation procedures
- Maryellen Giger, Emily Townley

# Medical Imaging has the largest number of FDA Cleared Algorithms

FDA Cleared Algorithms (N=950)



| Medical Specialty           | FDA devices |
|-----------------------------|-------------|
| Radiology                   | 723         |
| Cardiovascular              | 98          |
| Neurology                   | 34          |
| Hematology                  | 18          |
| Gastroenterology-Urology    | 14          |
| Ophthalmic                  | 10          |
| Anesthesiology              | 9           |
| Clinical Chemistry          | 8           |
| Pathology                   | 8           |
| General and Plastic Surgery | 6           |
| Microbiology                | 5           |
| Orthopedic                  | 5           |
| General Hospital            | 4           |
| Dental                      | 3           |
| Ear Nose & Throat           | 2           |
| Immunology                  | 1           |
| Obstetrics and Gynecology   | 1           |
| Total                       | 950         |



# AI algorithms in Radiology

- Large number of FDA approved methods due to
  - demonstrated ability of AI to analyze medical images
  - established digital workflows and universal DICOM standards for image storage
- Large amounts of imaging data are needed to properly develop, train and evaluate AI algorithms - hundreds of thousands or millions (Willemink et al *Radiology* 2020)
- There is a lack of accessible, curated and representative training data
- The majority of medical image data are stored in isolated PACS systems, which work well for the local clinical needs
- Clinical data are rarely shared externally due to a lack of motivation, a lack of resources, and privacy concerns
- Most research groups and industry have access limited to sets with small sample sizes from fixed geographic areas

# What went wrong with AI/ML methods using imaging?

MIT  
Technology  
Review

Featured Topics Newsletters Events Podcasts

Artificial intelligence / Machine learning

## Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

### nature machine intelligence

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature machine intelligence > analyses > article

Analysis | [Open Access](#) | Published: 15 March 2021

## Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts , Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

*Nature Machine Intelligence* **3**, 199–217 (2021) | [Cite this article](#)

### 1) Poor quality of data, “Frankenstein data sets”

- Misabeled data
- Multiple unknown sources
- Duplicate data (training and testing)
- No traceability, limited quality control
- Lack of external validation

### 2) Lack of communication between AI/ML experts and Medical/ Biomedical experts

- Lack of valid ground truth
- Incomplete understanding of independent testing (i.e. no sequestered data)

### 3) Representativeness

- Data collected for a specific clinical task
- Specific populations, don’t reflect real-world variations

# What are the large data sets?

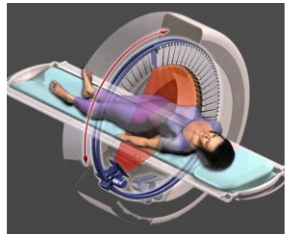
| Database                      | Access | Modalities  | # of Institutions | # of exams (pending) |
|-------------------------------|--------|-------------|-------------------|----------------------|
| MIDRC                         | Free   | Multiple    | Multiple          | 189,997 (377,400)    |
| Stanford AIMI                 | Free   | Multiple    | Single            | 285,182              |
| TCIA                          | Free   | Multiple    | Multiple          | 92,771               |
| ChestX-ray8                   | Free   | Radiography | Single            | 30,805               |
| National Lung Screening Trial | Free   | CT          | Multiple          | 26,254               |
| Proscia                       | Paid   | Multiple    | Multiple          | 1,000,000+           |
| Gradient Health               | Paid   | Multiple    | Multiple          | 1,000,000+           |
| RadImageNET                   | Paid   | Multiple    | Multiple          | 1,000,000+           |
| ARPA-H INDEX                  | Paid   | Multiple    | Multiple          | In development       |

In 2023, there were over 607,000,000 medical imaging procedures performed at over 12,000 imaging centers in the United States

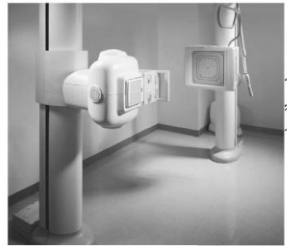
# *Medical Imaging Data Resource Center (MIDRC)*

- Established August 2020
- NIBIB Contract 75N92020D00021, ARPA-H Contract 75N92023F00002
- Created as an open, curated, diverse image data commons
- A partnership between the AAPM, ACR and RSNA, supported by NIBIB, hosted at University of Chicago, and on the Gen3 data platform
- Continued funding through ARPA-H and NAIRR

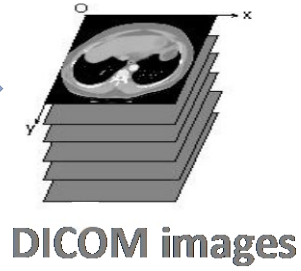
## Many Imaging centers



CT



XR



Deidentification



Curation

Annotation

Quality Assessment

Sequestration

Extraction of search data

Presentation of search data

## AI/ML algorithm developers

Cohort selection

Image download

Testing

Challenges

Guidelines

Metrics

**Total  
ingested into  
MIDRC**



# of Imaging Studies

**567,397**

**Undergoing  
MIDRC Data  
Quality and  
Harmonization**



# of Imaging Studies

**189,997**

**Released to  
the public by  
MIDRC**

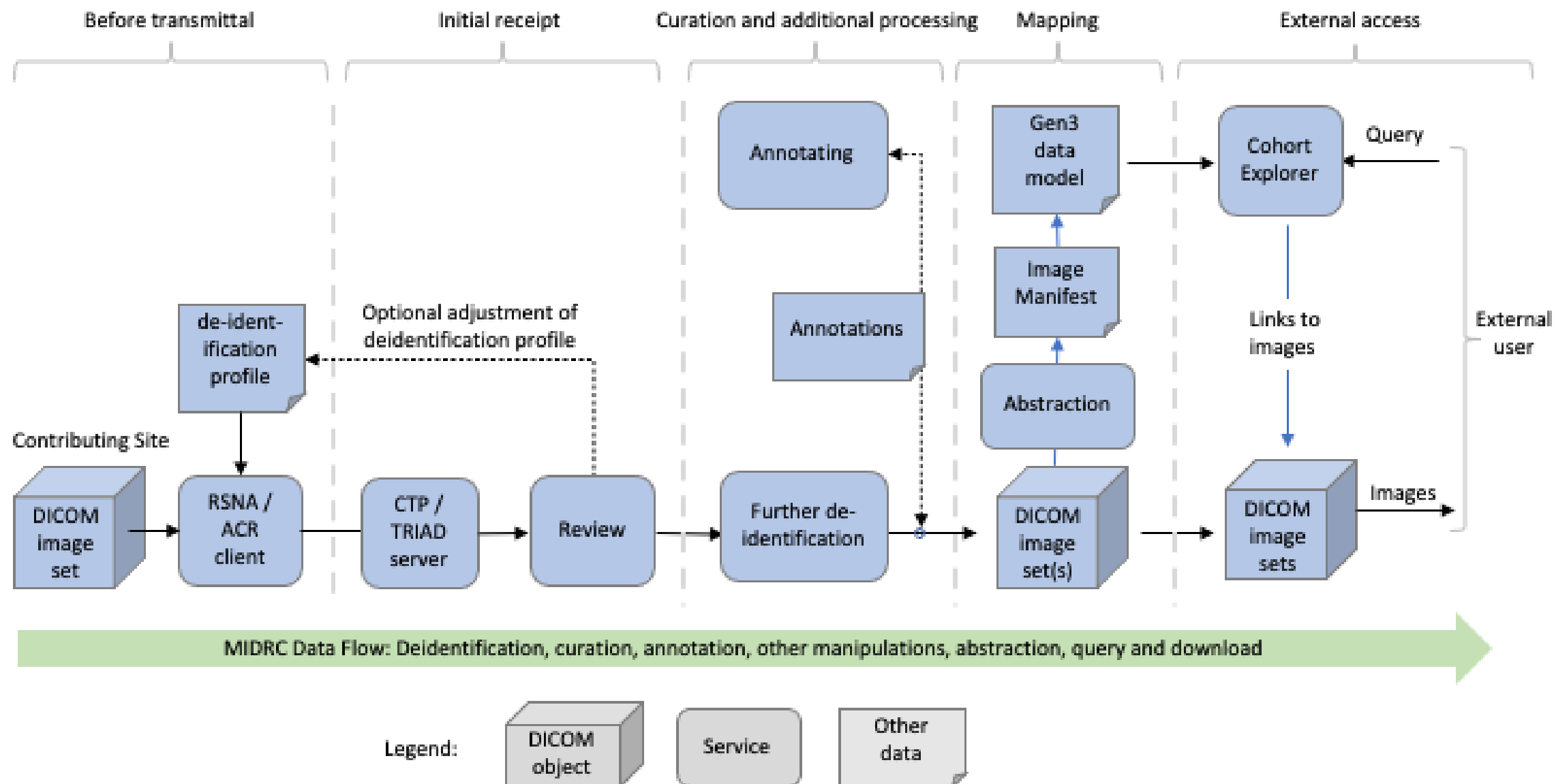


# of Imaging Studies

**377,400**

As of  
February 11, 2025

# MIDRC Data Flow



# MIDRC Data Explorer

[About](#) | [Get Started](#) | [Cite MIDRC](#) | [DUA](#) | [Resources](#) | [Gen3](#) | [Email Support](#) |

**Exploration**DictionaryQueryExample Analysis

**Cases** | **Annotations** | **Measurements** | **Imaging Studies** | **Data Files**

[Explorer Filters](#) | [Data Tools](#) | [Summary Statistics](#) | [Table of Records](#)

**Filters**

**Annotations**

**Imaging Properties**

**Study Properties**

[Collapse all](#)

**Case Demographics**

**LOINC Properties**

**Study UID**

Select...

**Study Modality**

☐ DX **82,947**

☐ CR **77,200**

☐ CT **23,952**

☐ MR **490**

☐ US **162**

9 more

**Study Description**

☐ XR CHEST 1 VIEW AP **21,911**

☐ XR CHEST 1 VW, FRONTAL **16,411**

☐ XR PORT CHEST 1V **12,428**

☐ no data **11,394**

☐ XR CHEST PORTABLE 1 VIEW **9,966**

1,884 more

**Age At Imaging**

**Download Table**

[Download File Manifest for Imaging Studies](#) . .

**Imaging Studies**

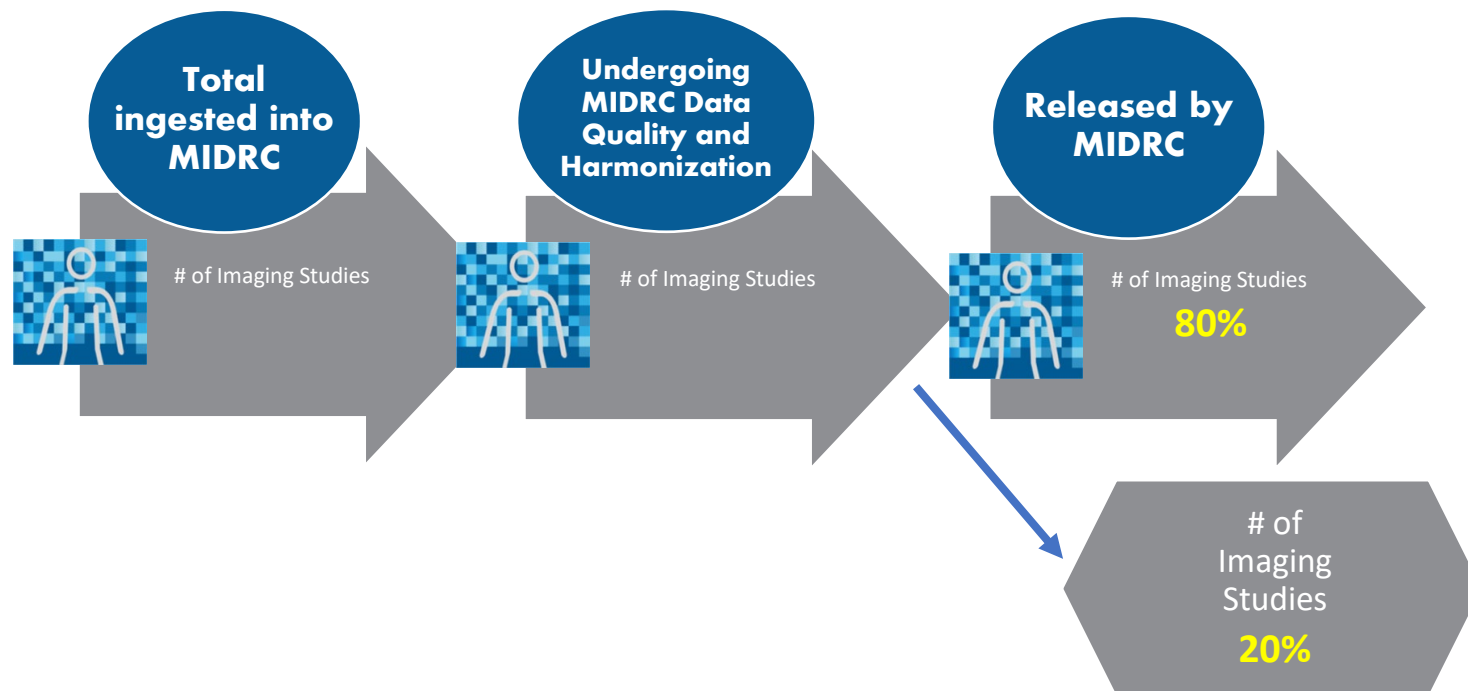
**184,483**

Showing 1 - 20 of 184,483 imaging\_studies

**Show Empty Columns**

|   | <b>Browse in DICOM viewer</b> | <b>Study Modality</b> | <b>Study Description</b>   | <b>Body Part Examined</b> | <b>Age At Imaging</b> | <b>Days To Study</b> | <b>Study Year</b> | <b>Study Year Shifted</b> | <b>Date</b> |
|---|-------------------------------|-----------------------|----------------------------|---------------------------|-----------------------|----------------------|-------------------|---------------------------|-------------|
| ▶ |                               | CR                    | XR CHEST 1 VW              | CHEST                     |                       | -275                 |                   | true                      | ACI         |
| ▶ |                               | DX                    | XR CHEST 1 VW, FRONTAL     | PORT CHEST                | 32                    | 0                    |                   | true                      | RSI         |
| ▶ |                               | CR                    | XR CHEST 1 VW PORTABLE     | CHEST                     |                       | 77                   |                   | true                      | ACI         |
| ▶ |                               | CR                    | CHEST PORT 1 VIEW (RAD)-CS | PORT CHEST                | 69                    | 231                  |                   | true                      | ACI         |
| ▶ |                               | DX                    | XR CHEST 1 VIEW AP         | CHEST                     | 85                    | 5                    |                   | true                      | RSI         |
| ▶ |                               | CR                    | XR CHEST PA AND LATERAL    |                           |                       | 136                  |                   | true                      | ACI         |
| ▶ |                               | CT                    | CT CHEST WO CONTRAST       |                           | 49                    | 6                    | 2020              | false                     | RSI         |

# MIDRC Sequestered Data Commons

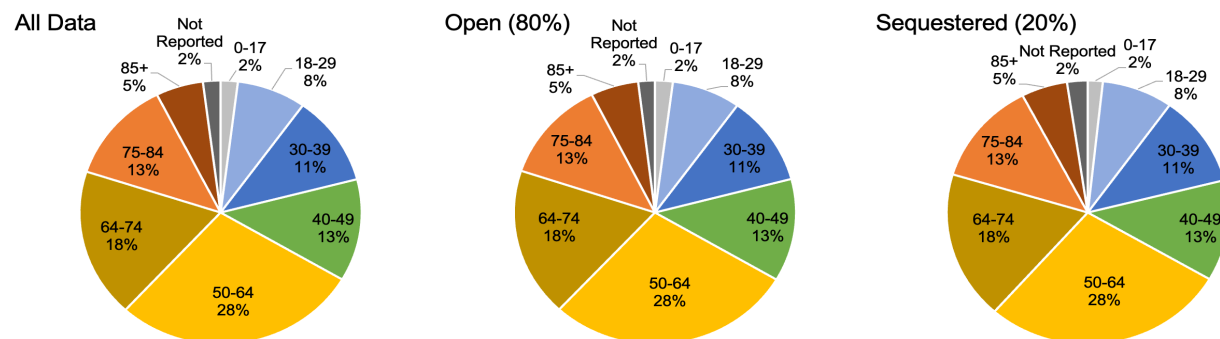


~80% are Publicly available

~20% are Sequestered Data

- Act as a large “test set” from which task-based samples can be drawn
- Accelerate translation of AI/ML algorithms
- Real-world performance of AI/ML on representative data

Balancing & checking demographic distributions

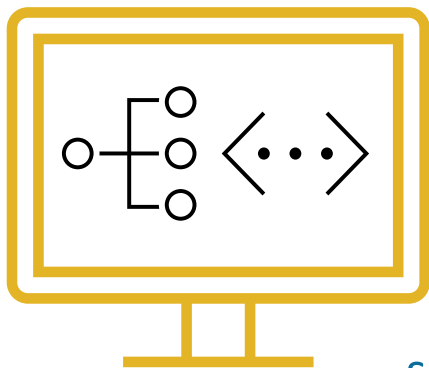


*Stratified sampling tool for sequestering, independent testing, and regulatory submissions*

|                                |                                                                                                                                 |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| MIDRC Tool Name                | MIDRC Stratified Sampling Algorithm                                                                                             |
| GitHub link or Web-facing link | <a href="https://github.com/MIDRC/Generalized_Stratified_Sampling">https://github.com/MIDRC/Generalized_Stratified_Sampling</a> |
| Publication                    | <a href="https://doi.org/10.1117/1.JMI.10.6.064501">https://doi.org/10.1117/1.JMI.10.6.064501</a>                               |

# Sequestered Commons for Real-World Evaluations and Translation through Regulatory to Clinical Care

Finalized user-developed algorithm for **specific clinical task** and **specific intended population**

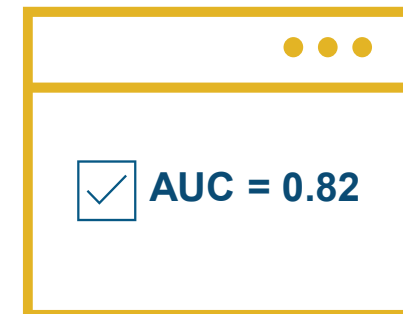


Testing request submitted to

Evaluation on the sequestered MIDRC testing data



MIDRC provides testing advice and independent performance assessment



MIDRC provided results

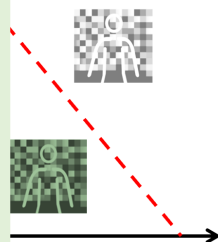
Report to developer



Report to FDA



- Created containerized workflows within Gen3 BRH (biomedical research hub; enclave) in a secure and isolated manner
- This overall evaluation process is a route for **sustainability**



# MIDRC BDF Imaging Hub (BIH) : Indexing Tool



Imaging Studies   Subject   **Imaging Series**

## Modality

| Name ▲                          | Imaging Series ▼ |
|---------------------------------|------------------|
| <input type="checkbox"/> CR     | 137,381          |
| <input type="checkbox"/> CT     | 406,384          |
| <input type="checkbox"/> DX     | 104,049          |
| <input type="checkbox"/> FUSION | 18               |
| <input type="checkbox"/> KO     | 40               |
| <input type="checkbox"/> M3D    | 2,328            |

+ 20 more

## Body Part Examined

| Name ▲                             | Imaging Series ▼ |
|------------------------------------|------------------|
| <input type="checkbox"/> ABD       | 2                |
| <input type="checkbox"/> ABD _ PEL | 9                |
| <input type="checkbox"/> ABD PEL   | 60               |
| <input type="checkbox"/> ABD PELV  | 4                |
| <input type="checkbox"/> Abdomen   | 2                |
| <input type="checkbox"/> ABDOMEN   | 17,923           |

+ 260 more

## Gender

| Name ▲                                | Imaging Series ▼ |
|---------------------------------------|------------------|
| <input type="checkbox"/> Female       | 433,269          |
| <input type="checkbox"/> Male         | 412,876          |
| <input type="checkbox"/> no data      | 17,545           |
| <input type="checkbox"/> Not Reported | 657,645          |
| <input type="checkbox"/> Unknown      | 279              |

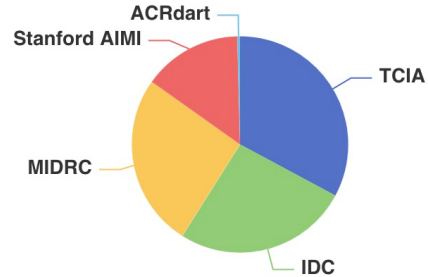
## Filters

No filters currently applied.

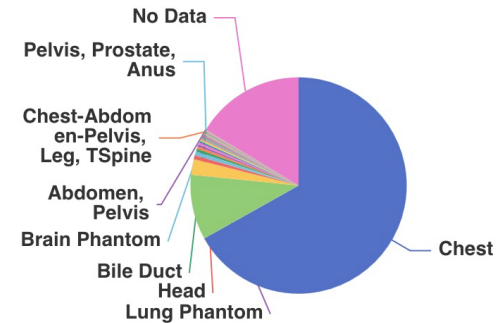
Download Table ▼

1,521,614 Imaging Series

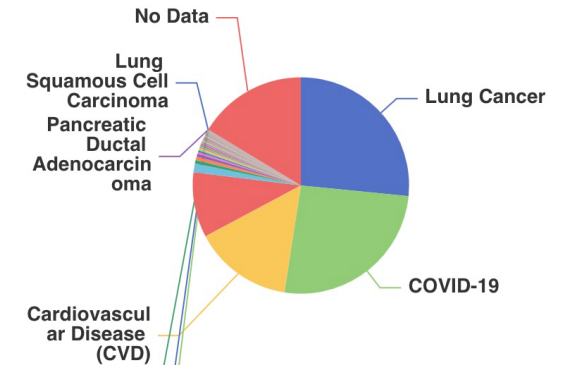
## Platform



## Primary Site



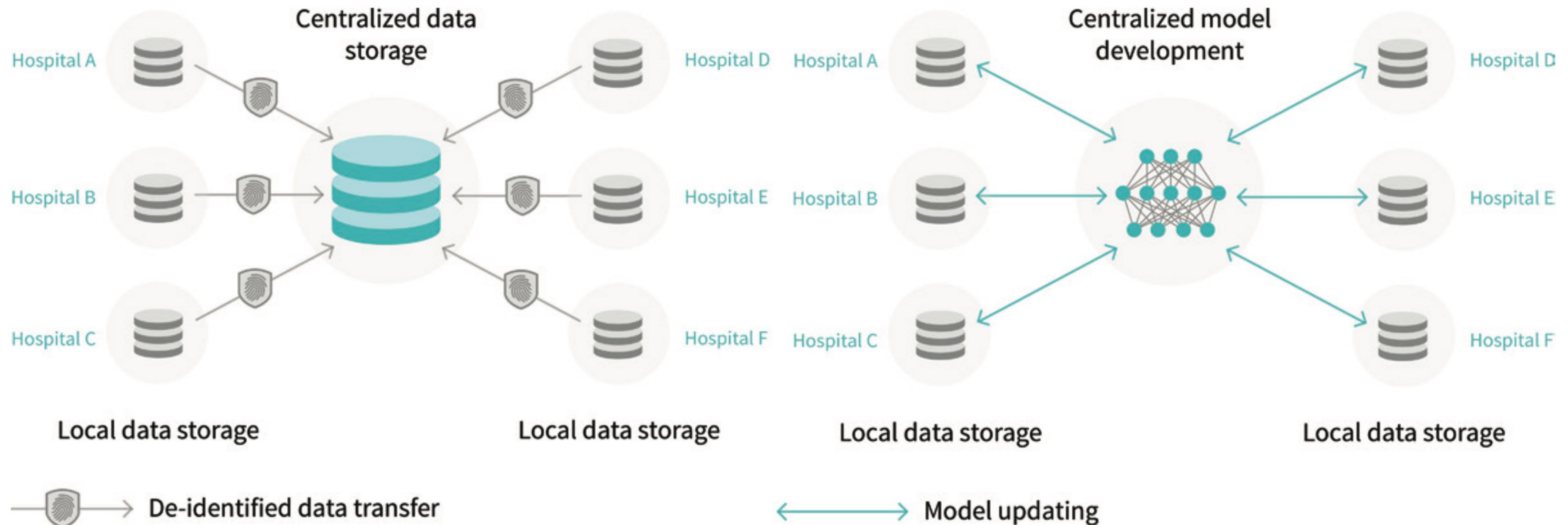
## Disease Type



| Modality | Body Part Examined | Gender | Primary Site | Disease Type                 | Platform                      | Collection      | Study Description              |
|----------|--------------------|--------|--------------|------------------------------|-------------------------------|-----------------|--------------------------------|
| CT       |                    | Male   | Lung         | COVID-19 (non-cancer)        | <a href="#">TCIA</a>          | COVID-19-NY-SBU | CT ABD AND PELVIS WITH IV CONT |
| XR       | CHEST              | Female | Chest        | cardiovascular disease (CVD) | <a href="#">Stanford AIMI</a> | CheXpert Plus   | XR CHEST Lateral               |

BIH is a federated hub that currently indexes 590,885 imaging studies

# Centralized versus federated data sharing



# Summary points

- “Numbers Are King, Quality Is Queen” - *Bob Gilles*
- Centralized repositories are valuable, but we need more data
- Options include centralized, indexed, federated, and hybrid combinations
- Curation is an essential aspect, which requires resources
- Sustainability of any approach is an essential component