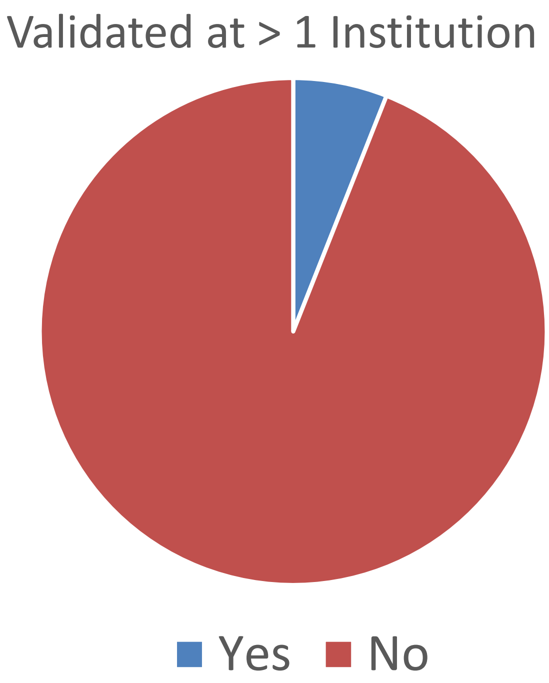
DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Independent Validation of AI Algorithms: Centralized and Distributed Solutions

Laura P Coombs, PhD
Vice President of Data Science and Informatics

Independent Validation

Externally Validated



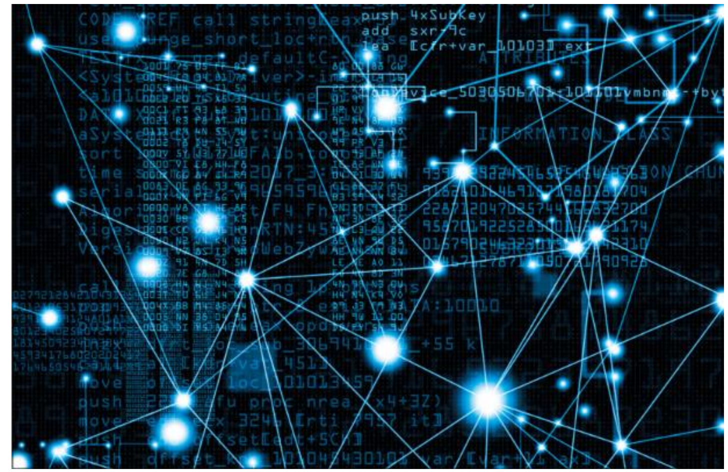
Evaluating Prediction Models

THE LANCET
Volume 393, Issue 10181, 20–26 April 2019, Pages 1577-1579

Comment

Reporting of artificial intelligence prediction models

Gary S Collins [a], Karel G M Moons [b]



"…artificial intelligence and machine learning prediction models must be appropriately developed, evaluated, and—if needed— tailored to different situations before they are used in daily medical practice…"

Bias

Ethical Tech / AI Ethics

AI is sending people to jail—a

Using historica
that machines

by **Karen Hao**

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 8 MONTHS AGO

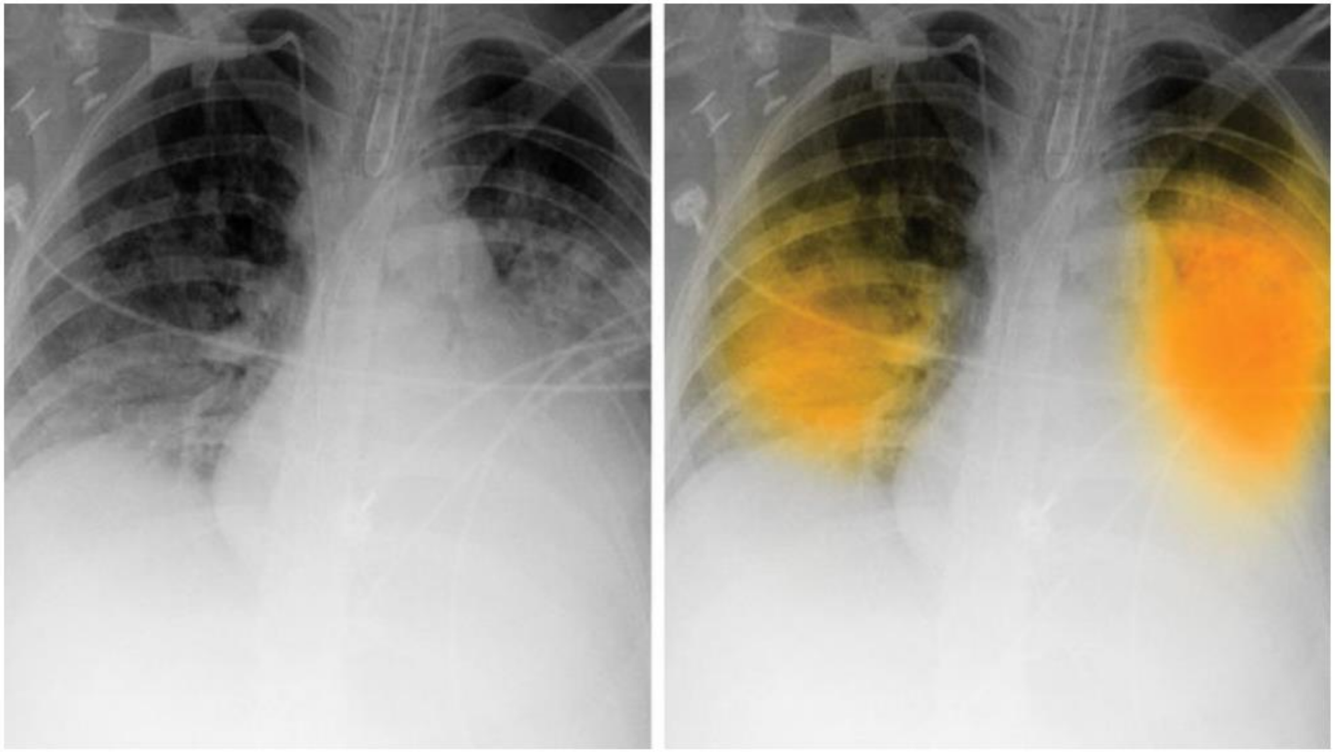
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY



Scientists are developing a multitude of artificial intelligence algorithms to help radiologists, like this one that lights up likely pneumonia in the lungs. ALBERT HSIAO AND BRIAN HURT/UC SAN DIEGO AIDA LABORATORY

Artificial intelligence could revolutionize medical care. But don't trust it to read your x-ray just yet

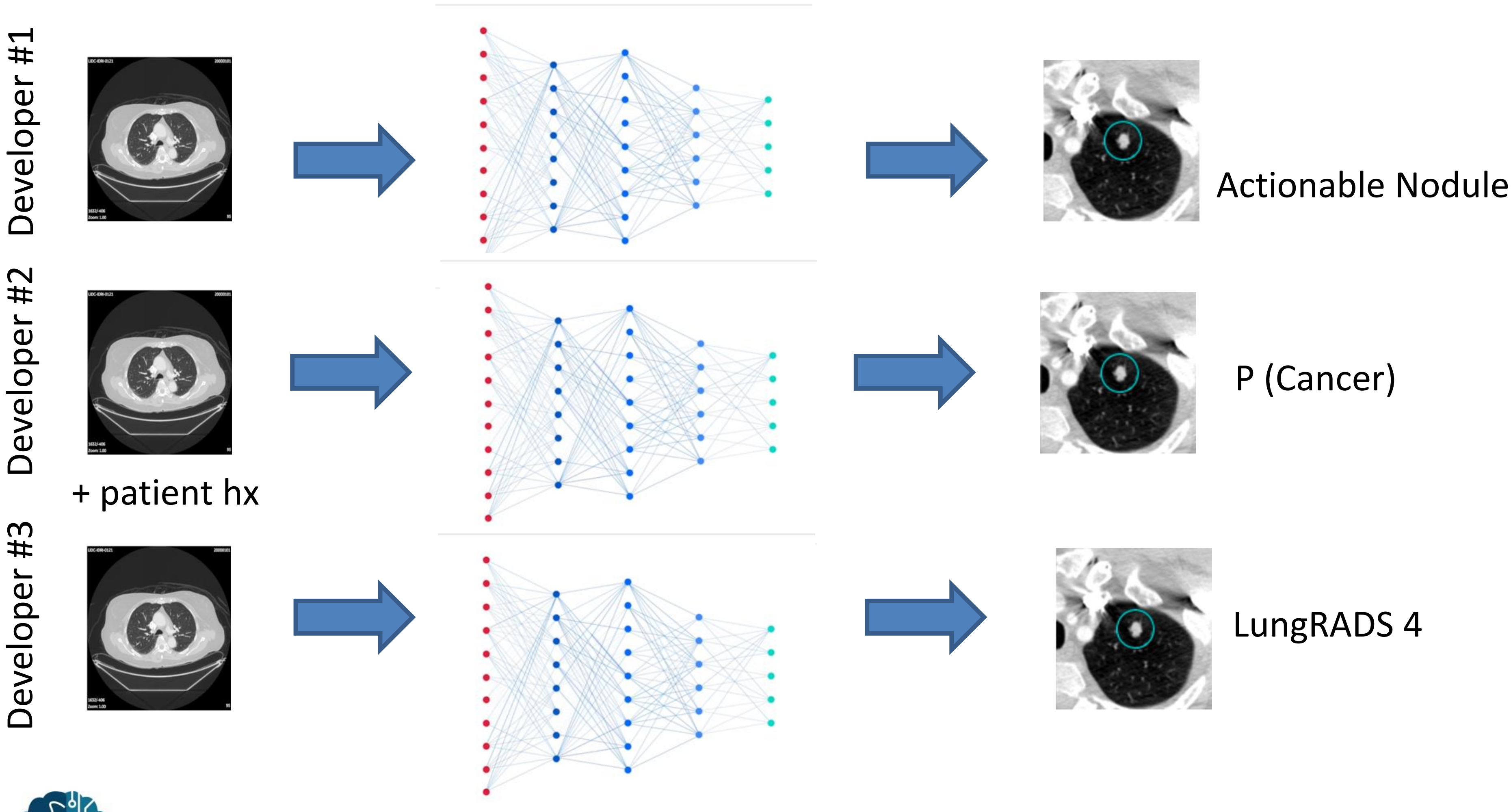
By Jennifer Couzin-Frankel | Jun. 17, 2019 , 12:45 PM

https://www.sciencemag.org/news/2019/06/artificial-intelligence-could-revolutionize-medical-care-don-t-trust-it-read-your-x-ray

Establish Standards and Certification Criteria

1. Establish common expectations for addressing specific clinical scenarios (e.g. BI-RADS)
2. Create well-qualified data sets that address explicit concerns about bias
3. Define standard performance metrics that establish a quality threshold
4. Validate models that address a specific clinical condition against these standards
5. Establish a controlled process for centralized and distributed validation

Lung Nodule Detection



Establish Standards and Certification Criteria

1. Establish common expectations for addressing specific clinical scenarios (e.g. BI-RADS)
2. Create well-qualified data sets that address explicit concerns about bias
3. Define standard performance metrics that establish a quality threshold
4. Validate models that address a specific clinical condition against these standards
5. Establish a controlled process for centralized and distributed validation

Characteristics that may affect accuracy

- **Scanner**: manufacturer, model, version
- **Acquisition parameters**: number of acquisitions, repetition time (TR), echo time (TE), and sampling bandwidth (SBW), pitch, detector configuration
- **Comorbidities**: diabetes, heart disease
- **Patient characteristics**: BMI, race, gender
- **Regional differences**: diet, environment

Establish Standards and Certification Criteria

1. Establish common expectations for addressing specific clinical scenarios (e.g. BI-RADS)
2. Create well-qualified data sets that address explicit concerns about bias
3. Define standard performance metrics that establish a quality threshold
4. Validate models that address a specific clinical condition against these standards
5. Establish a controlled process for centralized and distributed validation

Standard Performance Metrics

	Type of Primary Endpoint			
	Binary	Ordinal	Categorical	Continuous Variable
Example Use Cases	Pneumothorax, Pneumonia, Trauma fracture	Colon polyps	Appendicitis	Midline shift, Motor cortex, Scoliosis
Primary performance metrics	Sensitivity, specificity	Confusion matrix	Confusion matrix	Bias, repeatability (e.g. within-subject SD or CV), reproducibility
Primary statistical analyses	95% CIs for sensitivity, specificity	95% CIs for estimates of performance	95% CIs for estimates of performance	95% CIs of bias, repeatability, and reproducibility

The Appropriate Threshold is Context-Dependent

Algorithm	Examples	Metric
Classification	*RADS, Pneumothorax	Confusion matrix
Segmentation	Liver segmentation	DICE Coefficient
Estimation	Nodule Size, midline Shift	Bias, repeatability
Location	Nodule Detection	Dice Coefficient

Clinical Use	Risk
Prioritization in Work list	Low
Detection and Classification	Med
Diagnosis	High

Use Case	Certified Use (FDA)	Risk	Possible Result
Pneumothorax	Triage	Low	Pass
Pneumothorax	Diagnosis	High	Fail

DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Establish Standards and Certification Criteria

1. Establish common expectations for addressing specific clinical scenarios (e.g. BI-RADS)
2. Create well-qualified data sets that address explicit concerns about bias
3. Define standard performance metrics that establish a quality threshold
4. Validate models that address a specific clinical condition against these standards
5. Establish a controlled process for centralized and distributed validation

Certification Report

CAI-THOR00001 Pneumothorax Detection Certification Report

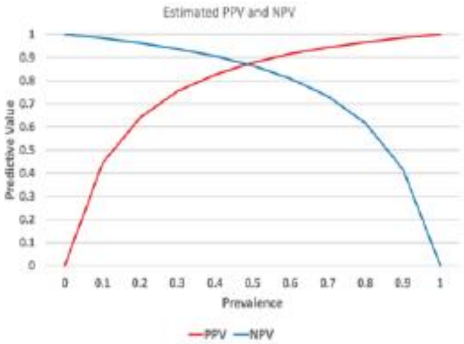


Figure 1: Estimated PPV (red) and NPV (blue) as a function of prevalence based on algorithm's point estimates from table 2.

ACR

DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Certify-AI

Table 3: Detection of Chest Tube

	Estimate	95% CI
Sensitivity* (n=452)	433/452 (95.8%)	[0.935, 0.973]
Specificity (n=1278)	1269/1278 (99.3%)	[0.987, 0.996]

*cases where the algorithm reported the chest tube as unknown are considered positive

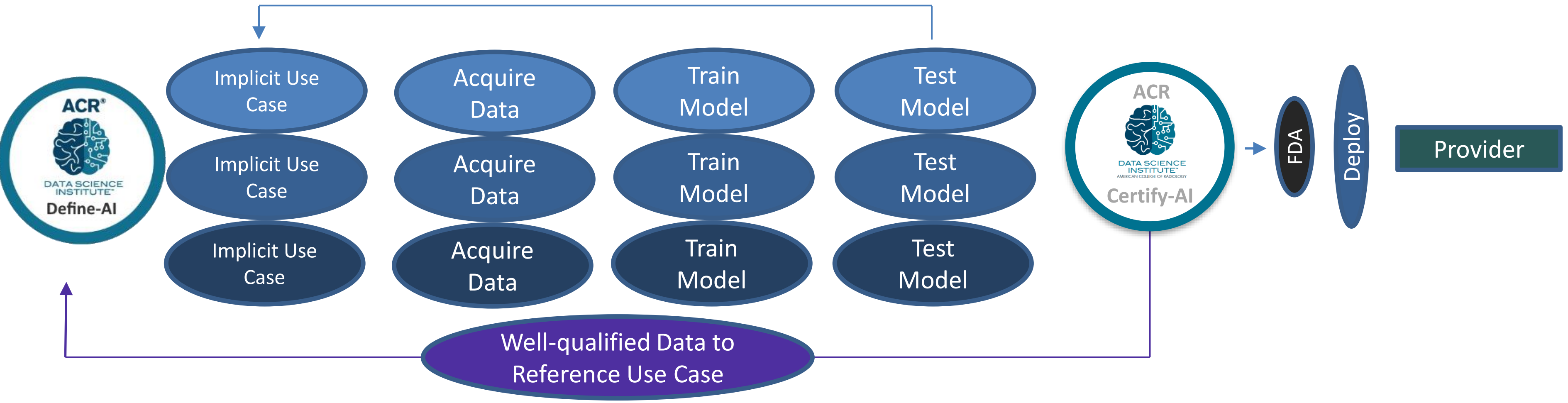
Conclusion: The algorithm demonstrates the ability to detect the presence of chest tubes with lower confidence bounds for sensitivity and specificity >0.90.

Table 4: Bias Assessment of Separation Measurements

Mean Bias (SE) n=852	-0.002 (0.07) [-0.143, 0.138]
Test that bias varies with magnitude of separation	p-value=0.676
Estimate of quadratic term	0.0007 (0.0005) [-0.0003, 0.0017]
Estimate of intercept (SE) [95% CI]	0.089 (0.22) [-0.34, 0.52]
Estimate of slope (SE) [95% CI]	0.983 (0.01) [0.983, 1.011]

DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Establish Standards & Certification Criteria

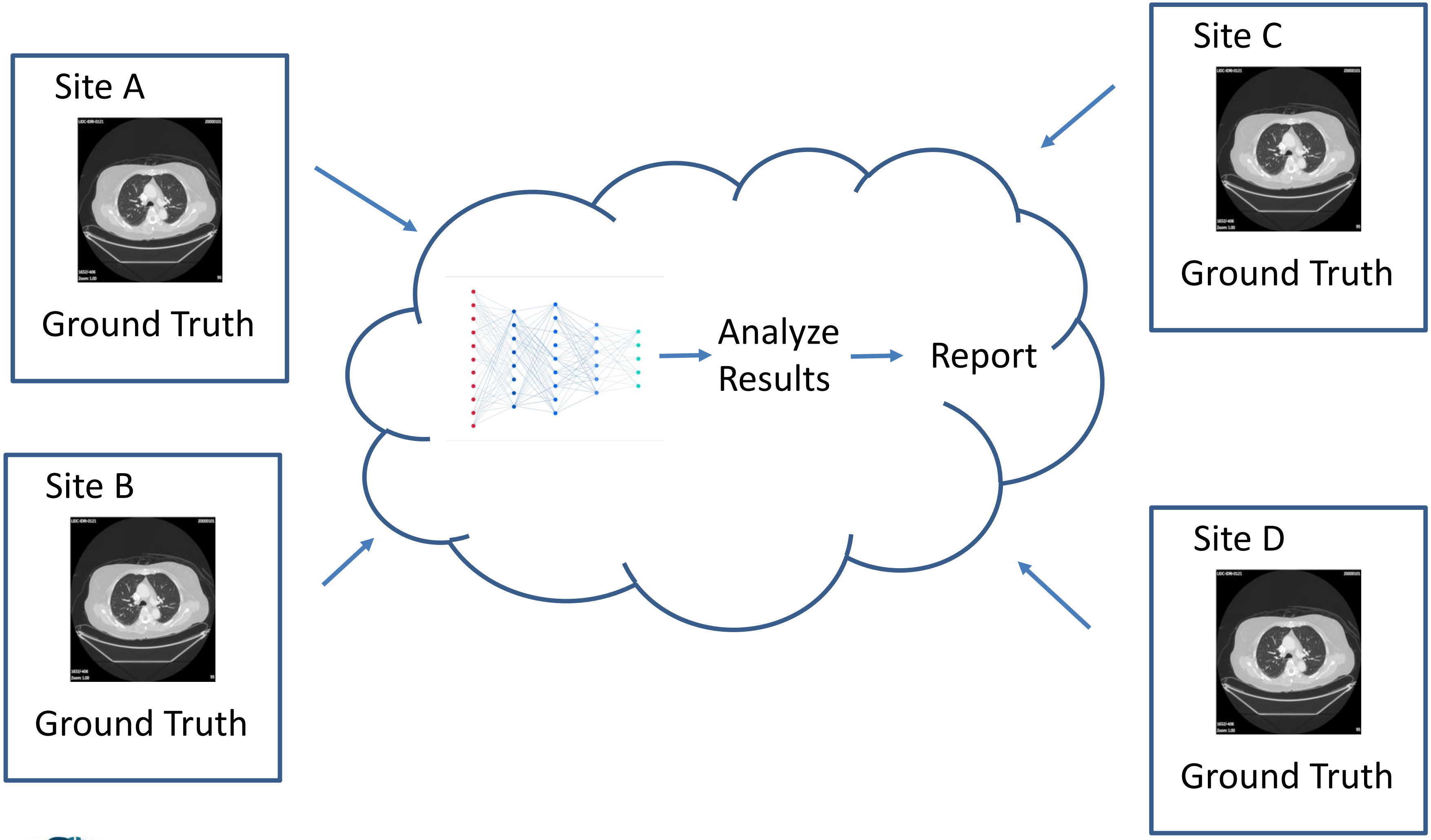


DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Establish Standards and Certification Criteria

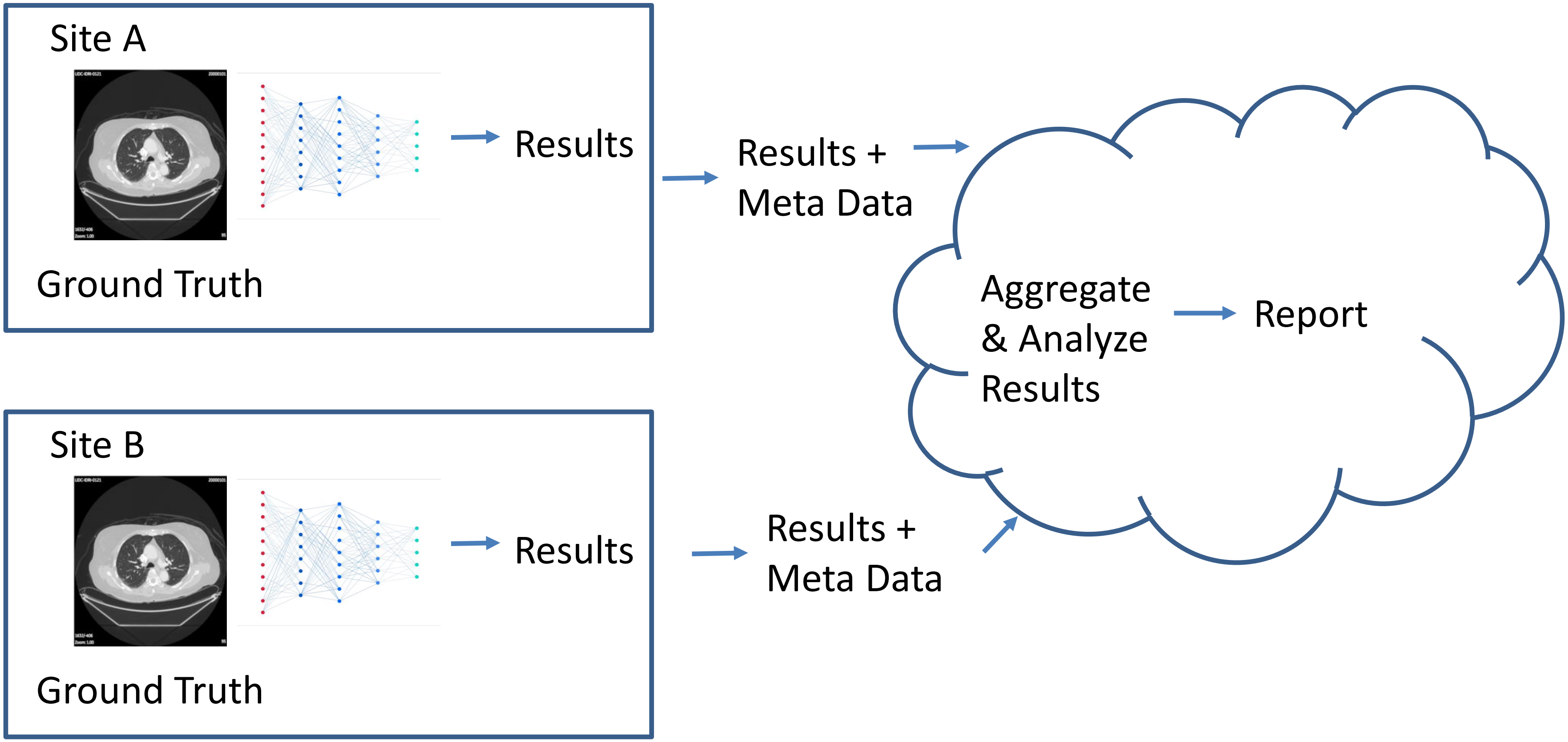
1. Establish common expectations for addressing specific clinical scenarios (e.g. BI-RADS)
2. Create well-qualified data sets that address explicit concerns about bias
3. Define standard performance metrics that establish a quality threshold
4. Validate models that address a specific clinical condition against these standards
5. Establish a controlled process for centralized and distributed validation

Central Validation



DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Distributed Validation

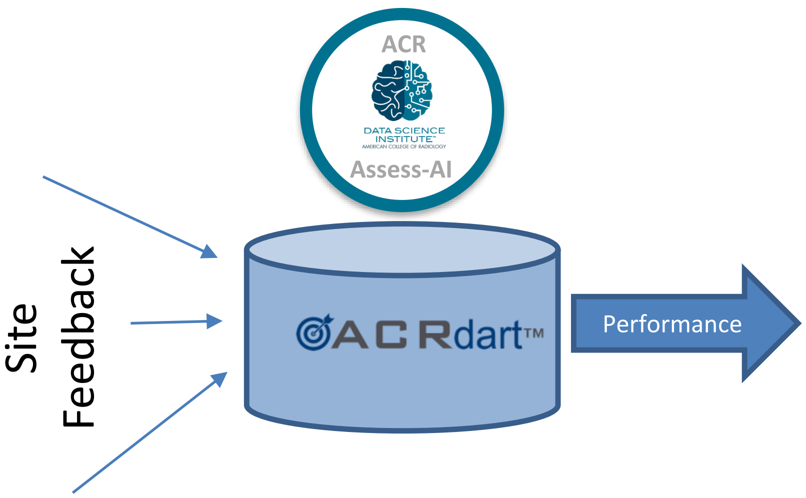


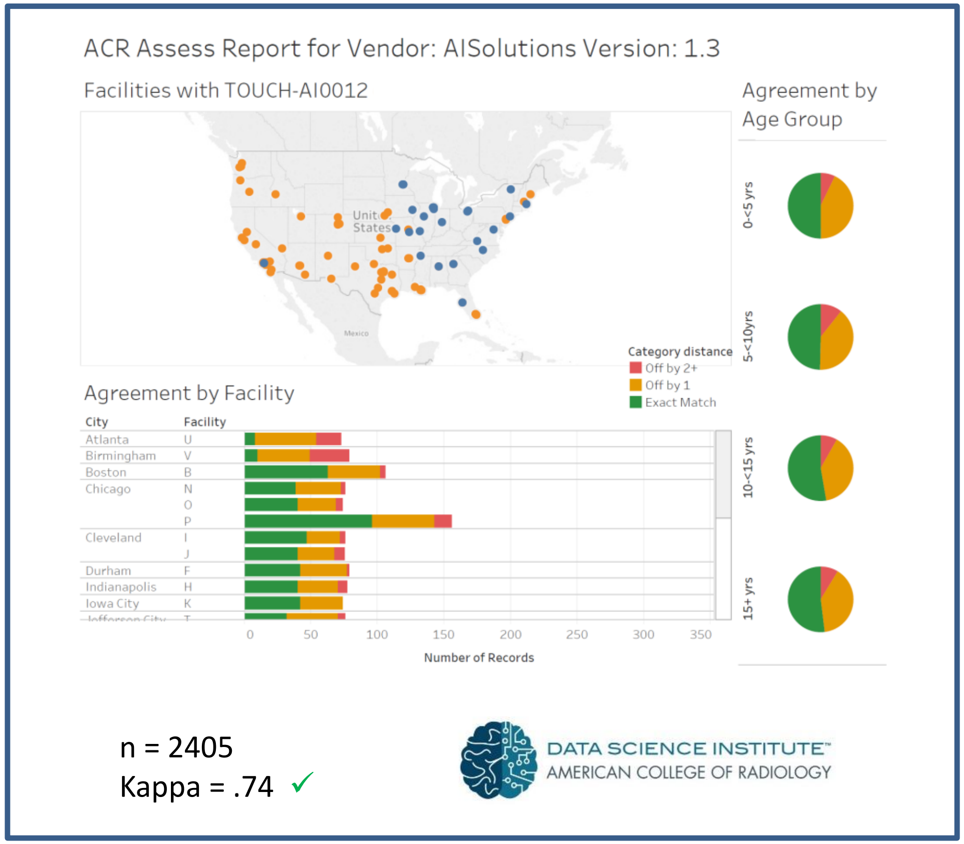
DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Distributed vs Centralized

- Issue with sharing data for centralized approach
- Issue with risk-adjustment for distributed
  - Either need to collect all metadata OR
  - Individual facility must provided info on incorrect cases

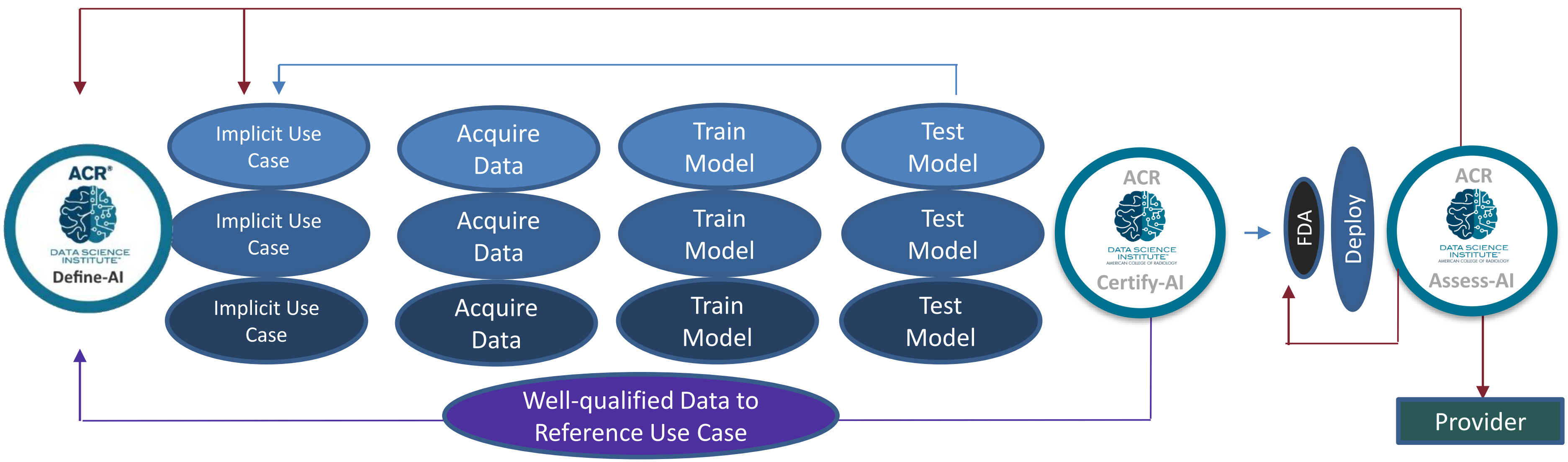
Monitoring and Benchmarking





DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Monitoring and Communication



DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

SUMMARY

- Standard inputs and outputs
- Well-qualified datasets
  - Central or Distributed
  - Diverse
- Standard methodology for evaluating algorithms
- Ability to monitor ongoing performance