Challenges of Creating, Validating, and Deploying AI Tools into Radiology Practice

Safwan S. Halabi MD - @SafwanHalabi

Associate Professor of Radiology,
Northwestern University Feinberg School of Medicine

Vice-Chair of Imaging Informatics,
Ann & Robert H. Lurie Children's Hospital of Chicago

Disclosures

Advisor	Bunker Hill Health
Consultant	Change Healthcare

Board Member, Society of Imaging Informatics in Medicine

Member, RSNA Informatics Committee

Chair, Data Science Standards Subcommittee

Motivations

Diagnostic errors play a role in up to 10% of patient deaths

21 percent of adults report having personally experienced a medical error

4% of radiology interpretations contain clinically significant errors

Improving Diagnosis in Health Care. National Academy of Medicine. Washington, DC: The National Academies Press, 2015.
Americans' Experiences with Medical Errors and Views on Patient Safety. Chicago, IL: University of Chicago and IHI/NPSF, 2017.
Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive Error in Radiology. *Am J Roentgenol*. 2016:1-11
Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades? *Am J Roentgenol*. 2007;188(5):1173-1178.

Motivations

Empower radiologists to provide high level diagnostic interpretation in setting of increased volume and limited resources

NOT to replace clinicians and radiologists

Radiologist disagreement

- Disagreement with colleagues – 25% of the time
- Disagreement with <u>themselves</u> – 30% of the time

Abujudeh, HH, Boland, GW, Kaewalai, R, et al. Abdominal and Pelvic Computed Tomography (CT) Interpretation: discrepancy rates among experienced radiologists. Eur Radiol.2010;20(8): 1952-7

ACTING AS AN EXPERT **CONSULTANT** TO YOUR REFERRING PHYSICIAN (THE DOCTOR WHO SENT YOU TO THE RADIOLOGY DEPARTMENT OR CLINIC FOR TESTING) BY AIDING HIM OR HER IN CHOOSING THE PROPER EXAMINATION, INTERPRETING THE RESULTING MEDICAL IMAGES, AND USING TEST RESULTS TO DIRECT YOUR CARE

TREATING DISEASES BY MEANS OF RADIATION (RADIATION ONCOLOGY) OR MINIMALLY INVASIVE, IMAGE-GUIDED THERAPEUTIC INTERVENTION (INTERVENTIONAL RADIOLOGY)

CORRELATING MEDICAL IMAGE FINDINGS WITH OTHER EXAMINATIONS AND TESTS

RECOMMENDING FURTHER APPROPRIATE EXAMINATIONS OR TREATMENTS WHEN NECESSARY AND CONFERRING WITH REFERRING PHYSICIANS

DIRECTING RADIOLOGIC TECHNOLOGISTS (PERSONNEL WHO OPERATE THE EQUIPMENT) IN THE PROPER PERFORMANCE OF QUALITY EXAMS

AJR FOCUS ON:

Implementing Machine Learning in Radiology Practice and Research

Marc Kohli[1]
Luciano M. Prevedello[2]
Ross W. Filice[3]
J. Raymond Geis[4]

OBJECTIVE. The purposes of this article are to describe concepts that radiologists should understand to evaluate machine learning projects, including common algorithms, supervised as opposed to unsupervised techniques, statistical pitfalls, and data considerations for training and evaluation, and to briefly describe ethical dilemmas and legal risk.

CONCLUSION. Machine learning includes a broad class of computer programs that improve with experience. The complexity of creating, training, and monitoring machine learning indicates that the success of the algorithms will require radiologist involvement for years to come, leading to engagement rather than replacement.

It is difficult to ignore the growing ... be straightforward to build, many topics out-

ELSEVIER
Canadian Association of Radiologists Journal xx (2018) 1–16
CANADIAN ASSOCIATION OF RADIOLOGISTS JOURNAL
www.carjonline.org

Health Policy and Practice / Santé: politique et pratique médicale

Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology

An Tang, MD, MSc[a,b,*], Roger Tam, PhD[c,d], Alexandre Cadrin-Chênevert, MD, BIng[e], Will Guest, MD, PhD[c], Jaron Chong, MD[f], Joseph Barfett, BESc, MSc, MD[g], Leonid Chepelev, MD PhD[h], Robyn Cairns, MSc, MD[i], J. Ross Mitchell, PhD[j], Mark D. Cicero, MD, BESc, FRCPC[g], Manuel Gaudreau Poudrette, MD[k], Jacob L. Jaremko, MD, PhD[l], Caroline Reinhold, MD, MSc[f], Benoit Gallix, MD[f], Bruce Gray, MD, FRCPC[g], Raym Geis, MD, FACR[m], for the Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group

[a]*Department of Radiology, Université de Montréal, Montréal, Québec, Canada*
[b]*Centre de recherche du Centre hospitalier de l'Université de Montréal, Montréal, Québec, Canada*
[c]*Department of Radiology, University of British Columbia, Vancouver, British Columbia, Canada*
[d]*School of Biomedical Engineering, University of British Columbia, Vancouver, British Columbia, Canada*
[e]*Department of Medical Imaging, CISSS Lanaudière, Université Laval, Joliette, Québec, Canada*
[f]*Department of Radiology, McGill University Health Center, Montréal, Québec, Canada*
[g]*Department of Medical Imaging, St. Michael's Hospital, University of Toronto, Toronto, Ontario, Canada*
[h]*Department of Radiology, University of Ottawa, Ottawa, Ontario, Canada*
[i]*Department of Radiology, British Columbia's Children's Hospital, University of British Columbia, Vancouver, British Columbia, Canada*
[j]*Department of Research, Mayo Clinic, Phoenix, Arizona, USA*
[k]*Department of Radiology, Université de Sherbrooke, Sherbrooke, Québec, Canada*
[l]*Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton, Alberta, Canada*
[m]*Department of Radiology, National Jewish Health, Denver, Colorado, USA*

Current Applications and Future Impact of Machine Learning in Radiology

Garry Choy, MD, MBA • Omid Khalilzadeh, MD, MPH[1] • Mark Michalski, MD • Synho Do, PhD • Anthony E. Samir, MD, MPH • Oleg S. Pianykh, PhD • J. Raymond Geis, MD • Pari V. Pandharipande, MD, MPH • James A. Brink, MD • Keith J. Dreyer, DO, PhD

From the Department of Radiology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit St, Boston, Mass 02114 (G.C., O.K., M.M., S.D., A.E.S., O.S.P., P.V.P., J.A.B., K.J.D.); and Department of Radiology, University of Colorado School of Medicine, Aurora, Colo (J.R.G.). Received August 16, 2017; revision requested October 3; final revision received January 2, 2018; accepted January 5. **Address correspondence to** G.C. (e-mail: *gchoy@partners.org*).

Current address:
[1]Department of Radiology, Mount Sinai Health System, Icahn School of Medicine at Mount Sinai, New York, NY.

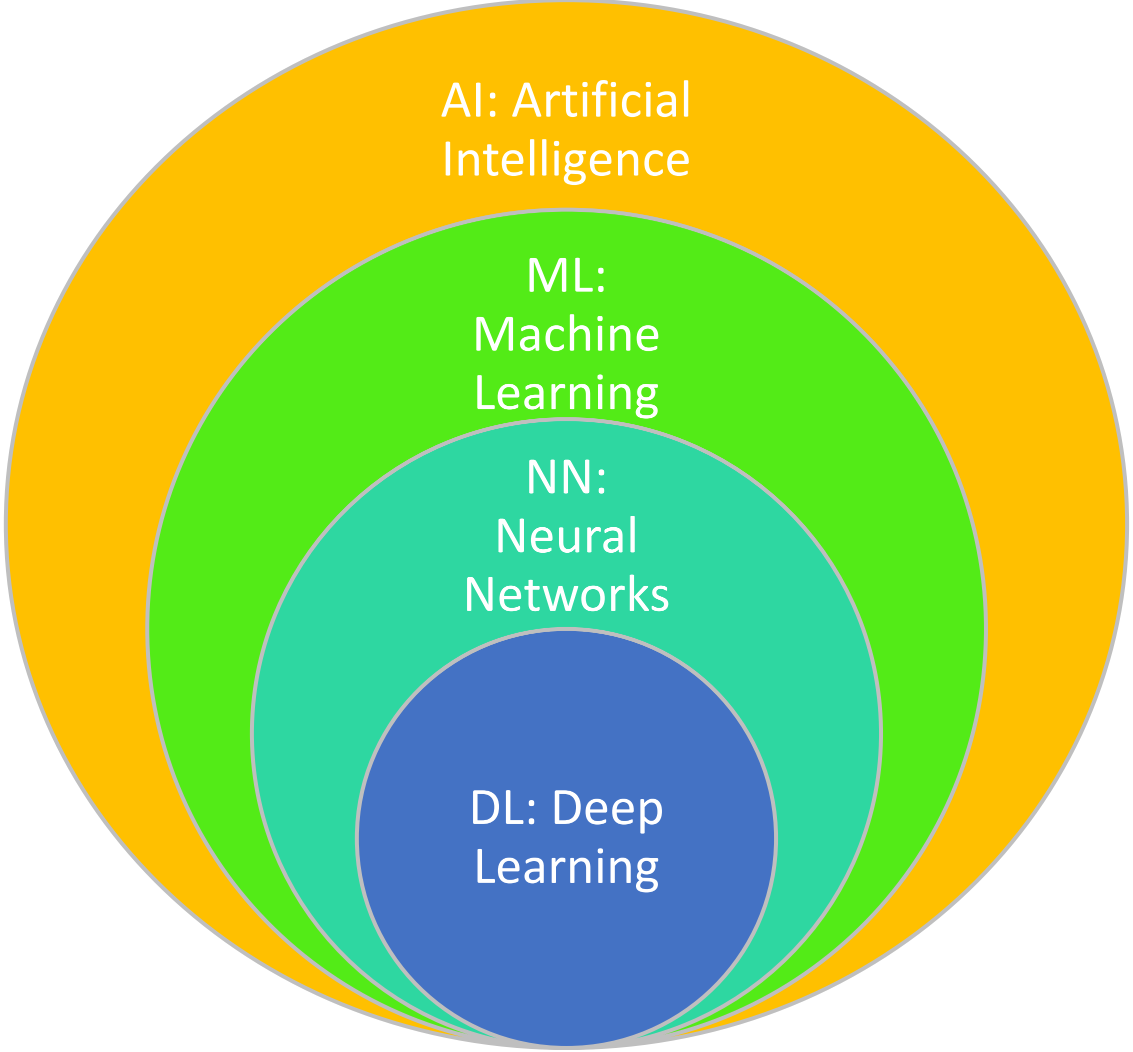
Conflicts of interest are listed at the end of this article.

Radiology 2018; 288:318–328 • https://doi.org/10.1148/radiol.2018171820 • Content code: IN

Recent advances and future perspectives of machine learning techniques offer promising applications in medical imaging. Machine learning has the potential to improve different steps of the radiology workflow including order scheduling and triage, clinical decision support systems, detection and interpretation of findings, postprocessing and dose estimation, examination quality control, and radiology reporting. In this article, the authors review examples of current applications of machine learning and artificial intelligence techniques in diagnostic radiology. In addition, the future impact and natural extension of these techniques in radiology practice are discussed.

Recent advances in machine learning offer promise in numerous industries and applications, including medical imaging (1). Within the innovations of data science, machine learning is a class of techniques and area of research

In supervised learning, data labels are provided to the algorithm in the training phase (there is supervision in training). The expected outputs are usually labeled by human experts and serve as ground truth for the algorithm. The goal



Definitions

- AI: When computers do things that make humans seem intelligent
- ML: Rapid automatic construction of algorithms from data
- NN: Powerful form of machine learning
- DL: Neural networks with many layers

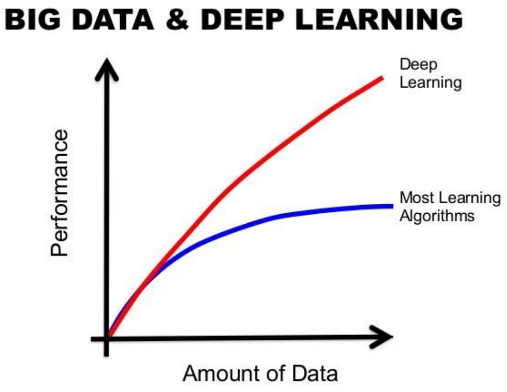
Augmented Intelligence

- Systems that are designed to **enhance** human capabilities
  - Contrasted with Artificial Intelligence, which is intended to replicate or replace human intelligence
- In healthcare (HC), a more appropriate term is 'augmented intelligence,' reflecting the **enhanced** capabilities of human clinical decision making when coupled with these computational methods and systems

AMA Passes First Policy Recommendations on Augmented Intelligence

For immediate release: Jun 14, 2018

Challenge #1: Dataset



- Collection of data
- Text and/or images

Data Challenges

- Do I have enough?
- Balanced?
- Representative?
- Annotated/labeled?
- De-identified?
  - Metadata
  - Facial scrubbing
  - Burned in data
- Sharing rights?

Attribute Name	Tag	Action	Comments
Station Name	(0008,1010)	Removed	Their values are only relevant to the equipment
Device Serial Number	(0018,1000)		
Institution Name	(0008,0080)	Removed	Their values are not normally relevant for research on image processing or aided diagnosis algorithms
Institution Address	(0008,0081)		
Referring Physician's Name	(0008,0090)		
Referring Physician's Address	(0008,0092)		
Referring Physician's Telephone Numbers	(0008,0094)		

Series Description	(0008,103E)		algorithms
Protocol Name	(0018,1030)		
Patient's Sex	(0010,0040)	Unchanged	Attributes that may be relevant for research algorithms
Patient's Size	(0010,1020)		
Patient's Weight	(0010,1030)		
Requested Procedure Description	(0032,1060)	Unchanged	Their values are important for image processing algorithms
Scheduled Procedure Step Description	(0040,0007)		
Performed Procedure Step Description	(0040,0254)		

AI Data Challenges

Heterogeneity of data

Heterogeneity of workflow

Determination of ground truth

Validation of AI models at different institutions

FDA approval of AI models for clinical use

Research Letter

September 22/29, 2020

Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms

Amit Kaushal, MD, PhD[1]; Russ Altman, MD, PhD[1]; Curt Langlotz, MD, PhD[2]

Author Affiliations | Article Information

JAMA. 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067

34 states aren't represented in any medical AI training sets

- Many academic research centers that do artificial intelligence and machine learning research are in health care hubs like Massachusetts, California, and New York.
- Data from California, home to Silicon Valley, was included in about 40% of the algorithms.

Table. US Patient Cohorts Used for Training Clinical Machine Learning Algorithms, by State[a]

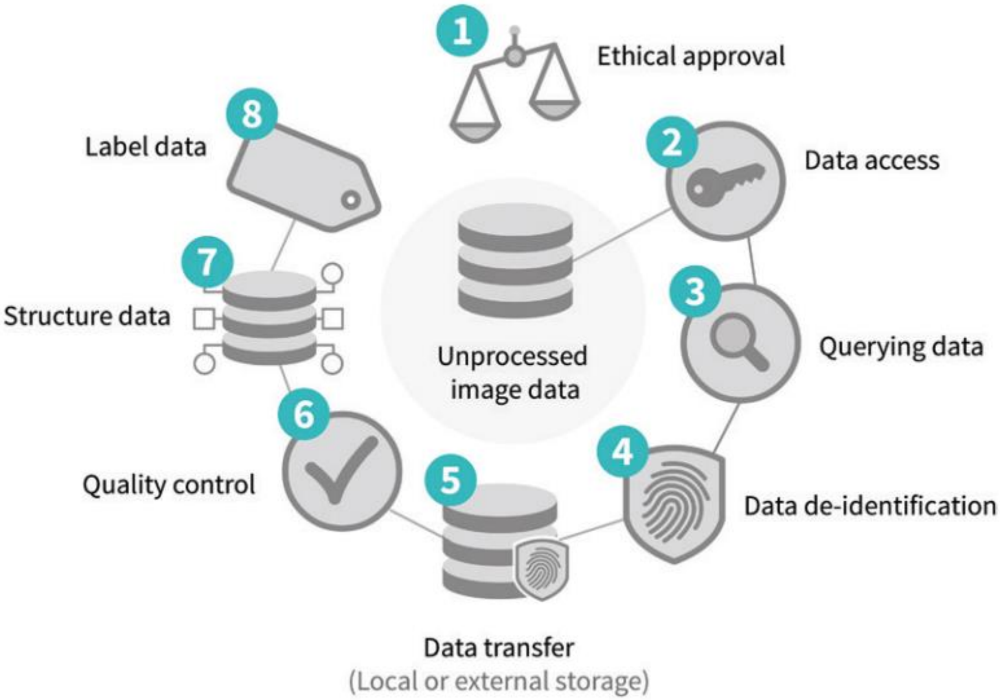
States	No. of studies
California	22
Massachusetts	15
New York	14
Pennsylvania	5
Maryland	4
Colorado	2
Connecticut	2
New Hampshire	2
North Carolina	2
Indiana	1
Michigan	1
Minnesota	1
Ohio	1
Texas	1
Vermont	1
Wisconsin	1

[a] Fifty-six studies used 1 or more geographically identifiable US patient cohorts in the training of their clinical machine learning algorithm. Thirty-four states were not represented in geographically identifiable cohorts: Alabama, Alaska, Arizona, Arkansas, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Iowa, Kansas, Kentucky, Louisiana, Maine, Mississippi, Missouri, Montana, Nebraska, Nevada, New Jersey, New Mexico, North Dakota, Oklahoma, Oregon, Rhode Island, South Carolina, South Dakota, Tennessee, Utah, Virginia, Washington, West Virginia, and Wyoming.

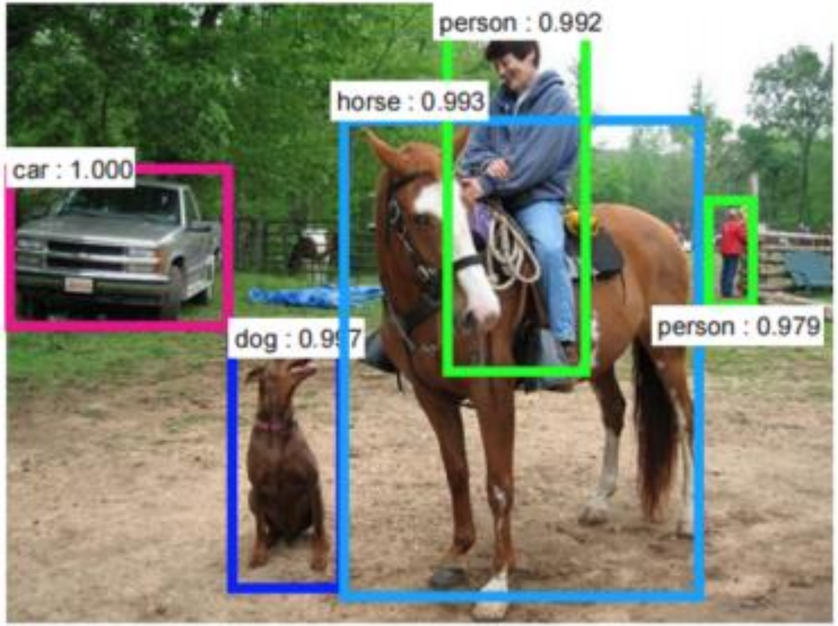
Handling Medical Image Data

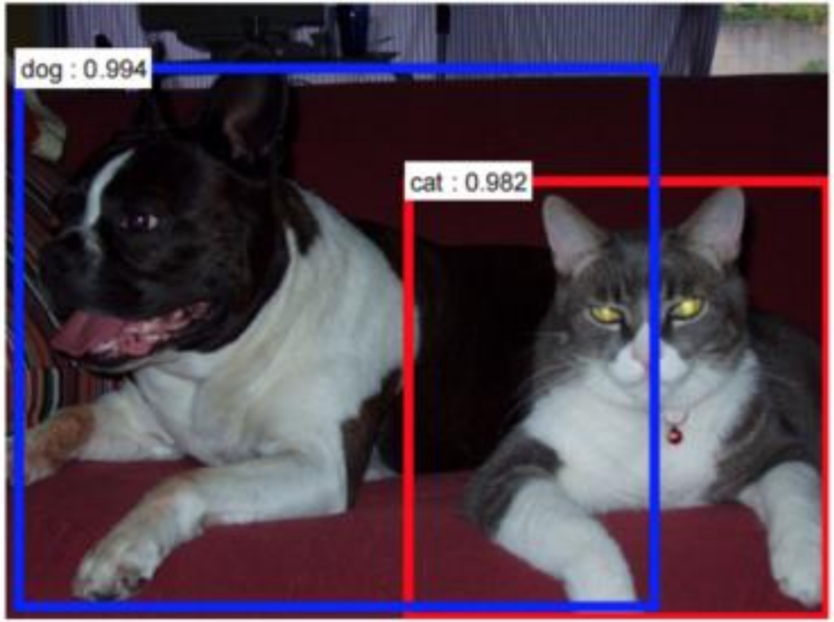
Willemink MJ. Published Online: February 18, 2020

https://doi.org/10.1148/radiol.2020192224

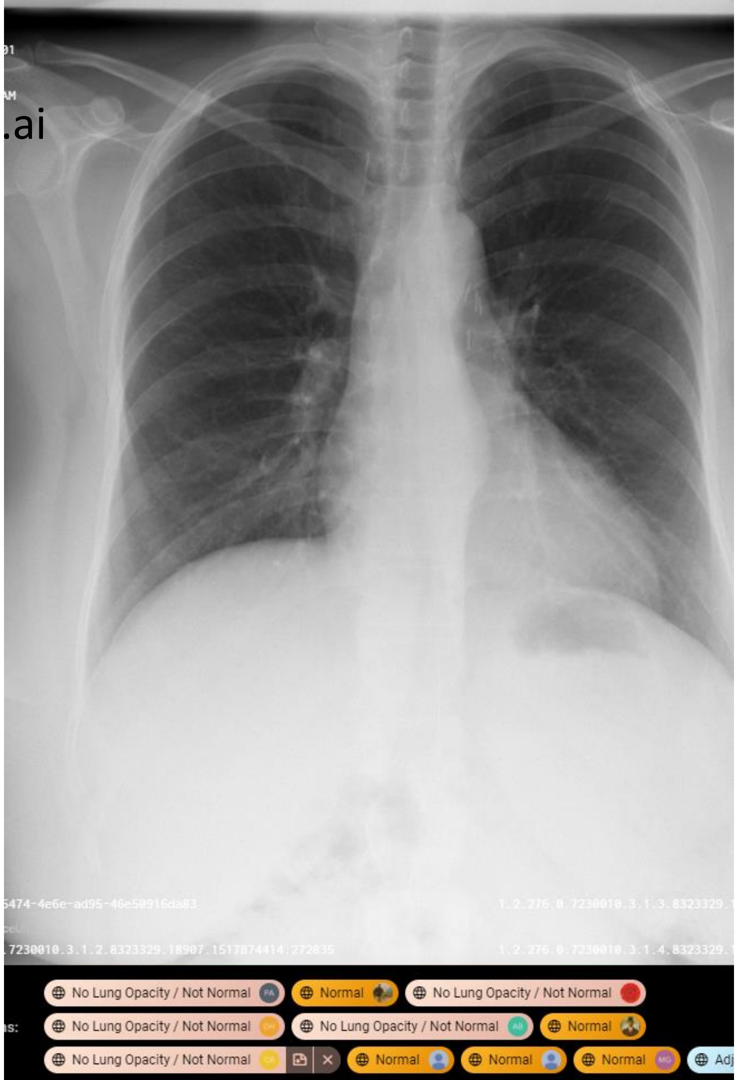


Challenge #2: Annotation

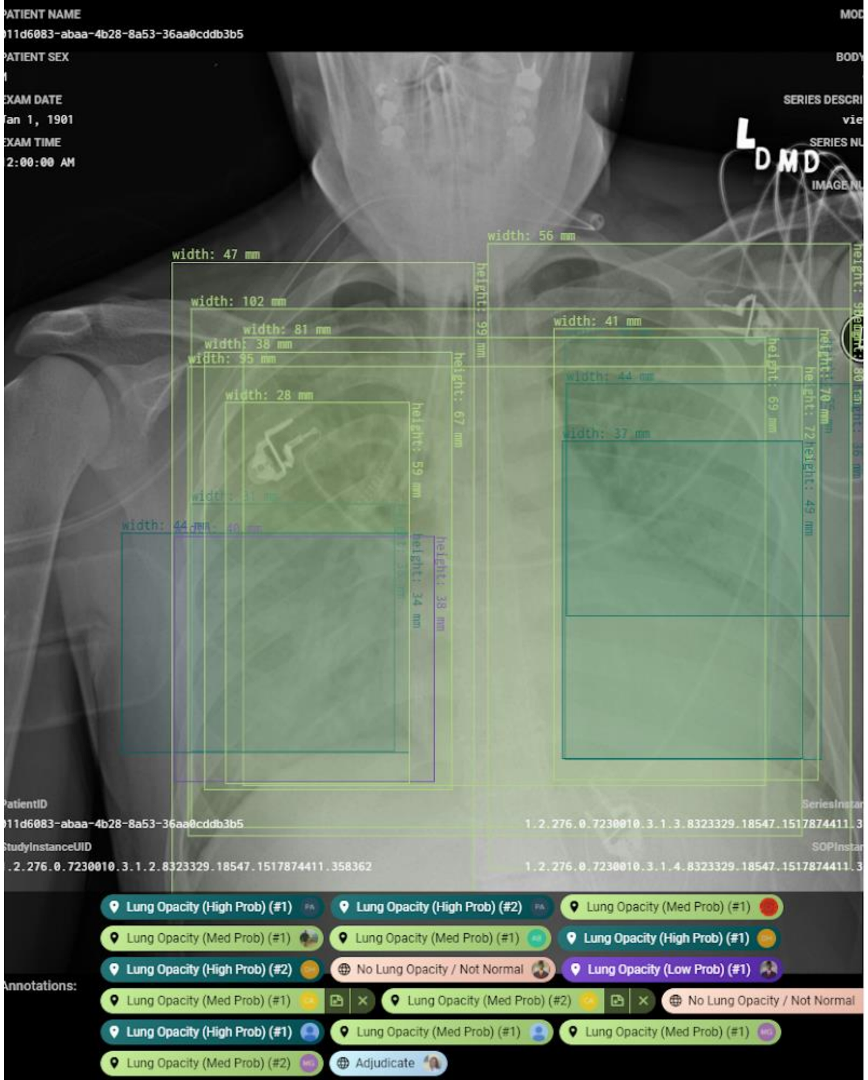




MD.ai



No Lung Opacity / Not Normal
Normal
No Lung Opacity / Not Normal
No Lung Opacity / Not Normal
No Lung Opacity / Not Normal
Normal
No Lung Opacity / Not Normal
Normal
Normal
Normal

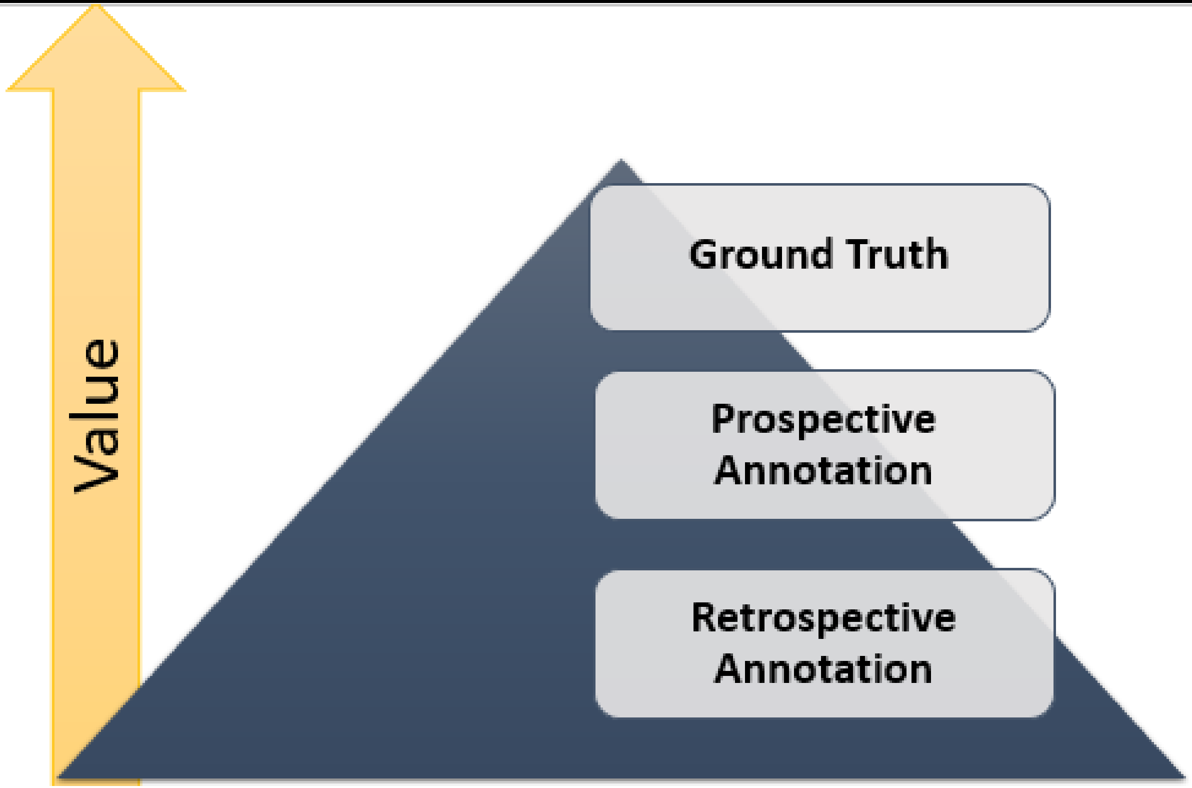


PATIENT NAME
011d6083-abaa-4b28-8a53-36aa0cddb3b5
PATIENT SEX
EXAM DATE
Jan 1, 1901
EXAM TIME
12:00:00 AM
width: 47 mm
width: 56 mm
width: 102 mm
width: 81 mm
width: 38 mm
width: 95 mm
width: 41 mm
width: 28 mm
width: 37 mm
height: 99 mm
height: 67 mm
height: 59 mm
height: 34 mm
height: 38 mm
PatientID
011d6083-abaa-4b28-8a53-36aa0cddb3b5
StudyInstanceUID
1.2.276.0.7230010.3.1.2.8323329.18547.1517874411.358362
1.2.276.0.7230010.3.1.3.8323329.18547.1517874411.3
1.2.276.0.7230010.3.1.4.8323329.18547.1517874411.3
Annotations:
Lung Opacity (High Prob) (#1)
Lung Opacity (High Prob) (#2)
Lung Opacity (Med Prob) (#1)
Lung Opacity (Med Prob) (#1)
Lung Opacity (Med Prob) (#1)
Lung Opacity (High Prob) (#1)
Lung Opacity (High Prob) (#2)
No Lung Opacity / Not Normal
Lung Opacity (Low Prob) (#1)
Lung Opacity (Med Prob) (#1)
Lung Opacity (Med Prob) (#2)
No Lung Opacity / Not Normal
Lung Opacity (High Prob) (#1)
Lung Opacity (Med Prob) (#1)
Lung Opacity (Med Prob) (#1)
Lung Opacity (Med Prob) (#2)
Adjudicate

A.I. Is Learning From Humans. Many Humans.

Artificial intelligence is being taught by thousands of office workers around the world. It is not exactly futuristic work.

Imaging Annotation Value



Algorithms

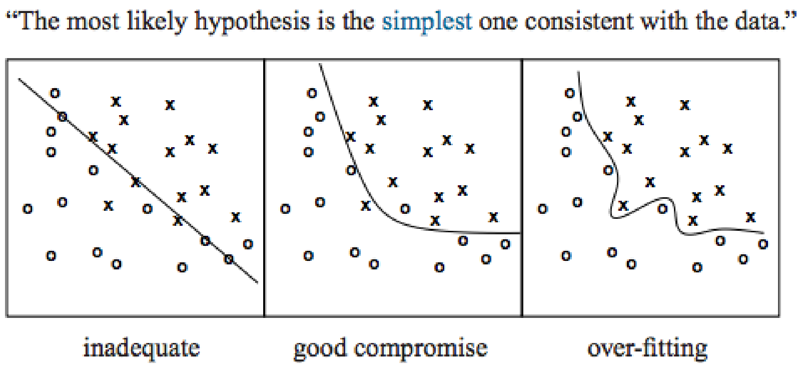
A set of rules or instructions given to an AI, neural network, or other machine to help it **learn on its own**

Clustering, classification, regression, and recommendations

Classification Models

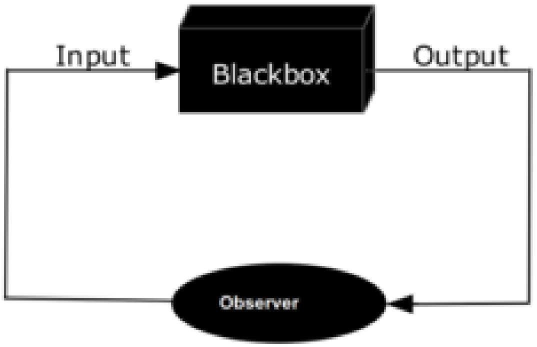
- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- Gradient-Boosted Tree
- Multilayer Perceptron
- Naive Bayes

Challenge #3: Validation



- Does the AI tool work in all scenarios?
  - Patient population
  - Imaging modalities
- Overfitting
  - The production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably
  - Overfitting and underfitting can occur in machine learning, in particular

Black Box



- In science, computing, and engineering, a black box is a device, system or object which can be viewed in terms of its inputs and outputs (or transfer characteristics), without any knowledge of its internal workings
- Its implementation is "opaque" (black)

Black Box Problem

Possible to observe incoming data (input) and outgoing data (output) in algorithmic systems, but their internal operations are not very well understood

Strong Black Boxes

- AI with decision-making processes that are entirely opaque to humans
- No way to determine (a) how the AI arrived at a decision or prediction, (b) what information is outcome determinative to the AI, or (c) to obtain a ranking of the variables processed by the AI in the order of their importance
- This form of black box cannot even be analyzed by reverse engineering the AI's outputs

Weak Black Boxes

- Decision-making process of a weak black box are also opaque to humans
- Weak black boxes can be reverse engineered or probed to determine a loose ranking of the importance of the variables the AI takes into account
- Allow a limited and imprecise ability to predict how the model will make its decisions

Algorithms

Although not all medical algorithms are black box, black box algorithms can allow the health system to leverage complex biological relationships well before those relationships are understood

Proof in the Pudding?

What if, inevitably, such an algorithm proves to be unreasonably effective at diagnosing a medical condition or prescribing a treatment, but you have no scientific understanding of how this link actually works?

Implementing AI in Radiology: Challenges

- Heterogeneity of data
- Heterogeneity of workflow
- Determination of ground truth
- Validation of AI models at different institutions
- FDA approval of AI models for clinical use

Radiology: Artificial Intelligence

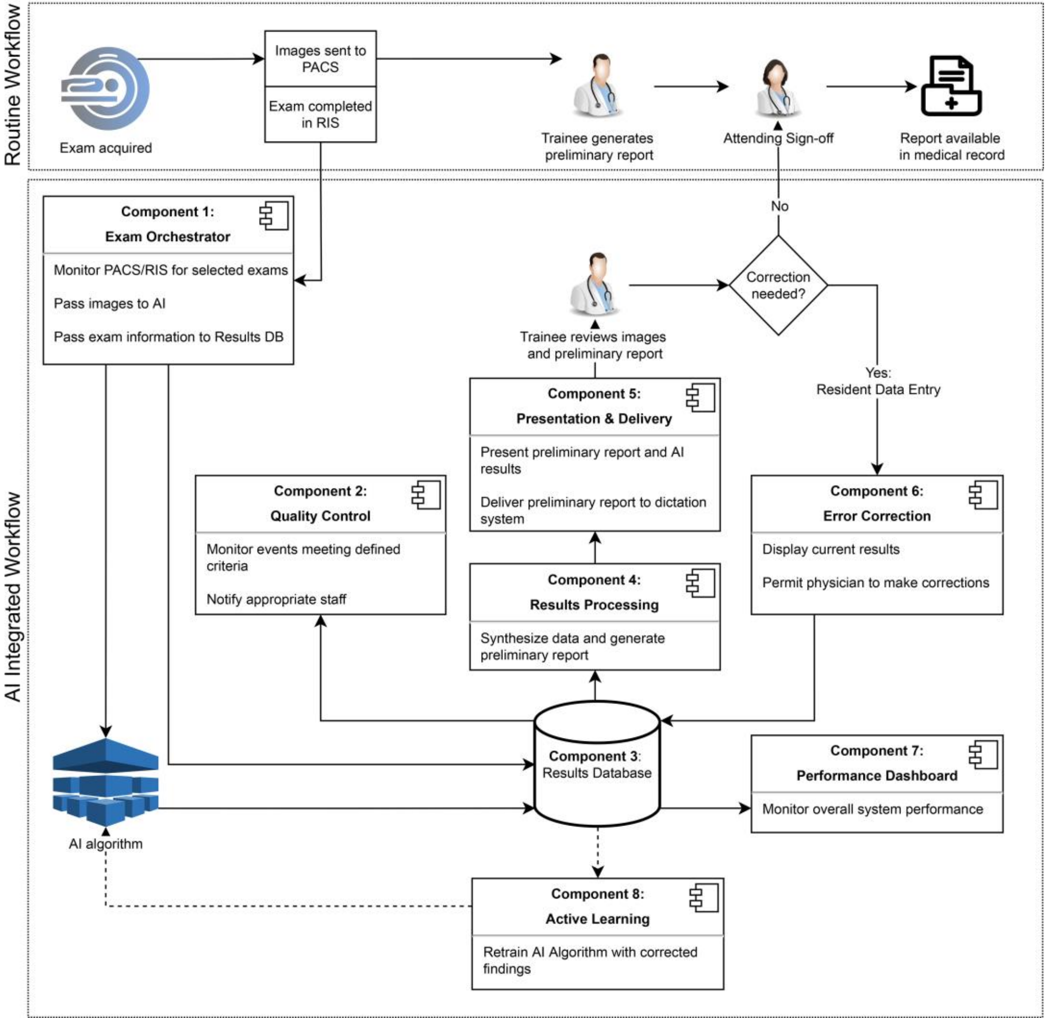
SPECIAL REPORT

Integrating AI Algorithms into the Clinical Workflow

Krishna Juluru, MD • Hao-Hsin Shih, MS • Krishna Nand Keshava Murthy, MS • Pierre Elnajjar, MS • Amin El-Rowmeim, MS • Christopher Roth, MD • Brad Genereaux, BS • Josef Fox, MD • Eliot Siegel, MD • Daniel L. Rubin, MD

From the Department of Radiology, Memorial Sloan-Kettering Cancer Center, 1275 York Ave, Box 29, New York, NY 10065 (K.J., H.H.S., K.N.K.M., P.E., A.E.R., J.F.); Department of Radiology, Duke University Medical Center, Durham, NC (C.R.); NVIDIA, Santa Clara, Calif (B.G.); Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, Md (E.S.); and Department of Radiology, Stanford University, Stanford, Calif (D.L.R.). Received January 12, 2021; revision requested March 3; revision received June 16; accepted July 14. **Address correspondence to** K.J. (e-mail: *juluruk@mskcc.org*).

Integrating AI algorithms into Clincal Workflow



Implementing AI: 3 Possible scenarios

1. AI on demand
2. Automated image analysis
3. Discrepancy management

P. Lakhani, NIBIB AI in Medical Imaging Workshop, Aug 23, 2018
P. Lakhani et al. JACR https://doi.org/10.1016/j.jacr.2017.09.044

Scenario 1

1. **AI on demand**
    - For a single image or series of images
    - PACS ➔radiologist ➔ AI server ➔ PACS, RIS, EHR
    - Radiologist would be in control of asking relevant AI interpretations
    - Requires manual step

Scenario 2

2. Automated AI image analysis

- Exams automatically sent to AI server (before reading)
- modality ➔ AI server ➔ PACS ➔radiologist ➔ RIS, EHR
- Helps to prioritizing reading order -> reduce TAT
- Radiologist views AI findings <u>before final report is made</u>
- Radiologist is able to ensure accuracy

Scenario 3

3. Discrepancy management

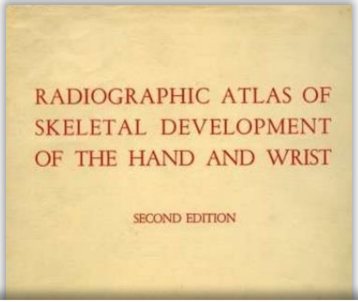
- As in 2. but results are <u>automatically</u> routed to RIS or EHR
- Requires <u>discrepancy management</u>
- AI -> preliminary -> RIS/EHR -> staff radiologist -> final
- Accurate AI needed (highly sens and spec), high confidence
- Fastest TAT although potential risk
- Might increase calls to radiology reading room
- Might have medicolegal consequences

Source: P. Lakhani, NIBIB AI in Medical Imaging Workshop, Aug 23, 2018

Radiology

Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs[1]

David B. Larson, MD, MBA
Matthew C. Chen, MS
Matthew P. Lungren, MD, MPH
Safwan S. Halabi, MD
Nicholas V. Stence, MD
Curtis P. Langlotz, MD, PhD



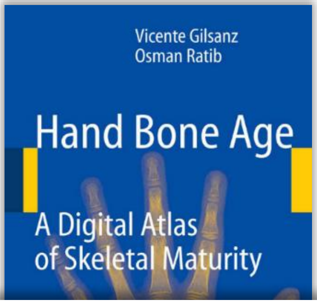


Table 2

Summary Statistics of Paired Interobserver Difference between Bone Age Estimate of Each Reviewer and Mean of the Other Three Human Reviewers' Estimates, Compared with That of Model

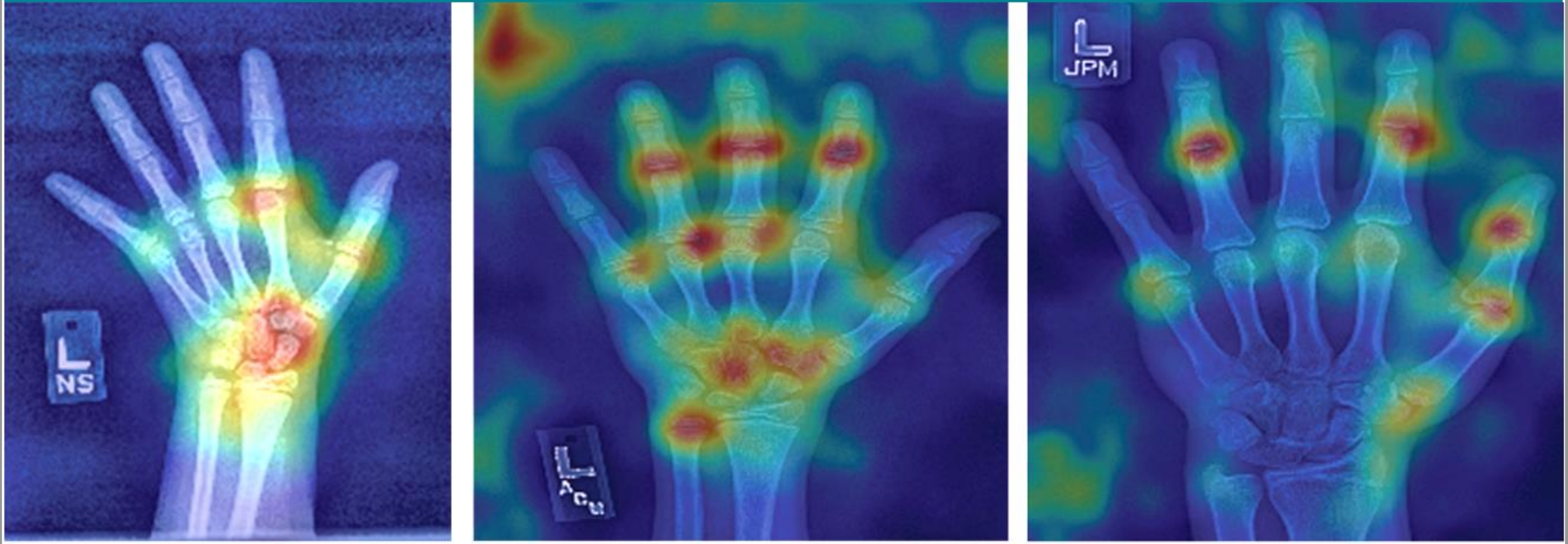
Variable	Clinical Report	Reviewer 1	Reviewer 2	Reviewer 3	Mean
MAD					
Reviewer	0.65	0.55	0.53	0.69	0.61
Model	0.51	0.53	0.53	0.53	0.52
P value (paired *t* test)	<.01	.50	.99	<.01	

Note.—Unless otherwise noted, data are expressed as years. The authors of the clinical report were treated collectively as a single reviewer.

https://doi.org/10.1148/radiol.2017170236

Saliency Maps

Figure 6



a. b. c.

Figure 6: Original image with superimposed saliency map for sample hand radiographic images in three male patients age 4 years (**a**), 15 years (**b**), and 17 years (**c**).

Saliency Maps

Original Research

Assessing the (Un)Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging

Nishanth Arun*, Nathan Gaw*, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, Jayashree Kalpathy-Cramer See fewer authors

* N.A. and N.G. contributed equally to this work.

Author Affiliations

Published Online: Oct 6 2021 | https://doi.org/10.1148/ryai.2021200267

Saliency Maps

- Using two large publicly available radiology datasets (SIIM-ACR Pneumothorax Segmentation and RSNA Pneumonia Detection), quantified the performance of eight commonly used saliency map techniques in regard to their 1) localization utility (segmentation and detection), 2) sensitivity to model weight randomization, 3) repeatability, and 4) reproducibility. We compared their performances versus baseline methods and localization network architectures, using area under the precision-recall curve (AUPRC) and structural similarity index (SSIM) as metrics.

Saliency Maps

- The use of saliency maps in the high-risk domain of medical imaging warrants additional scrutiny and recommend that detection or segmentation models be used if localization is the desired output of the network.

Implementing BA Model Clinically

- Institutional Review Board (IRB)
- Data Use Agreement (DUA)
- Consent (Patient? Radiologist?)
- Interfaces
- Workflow
- AI Model

Validation of BA tool by Randomized Control Trial

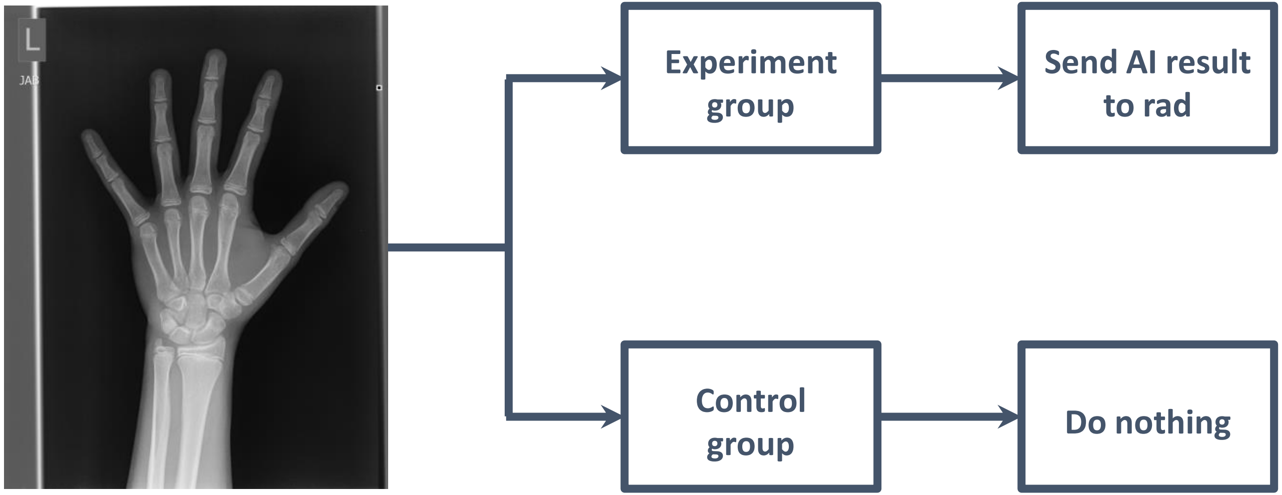
How does exposing the prediction of the AI model to the attending radiologist prospectively affect diagnosis?

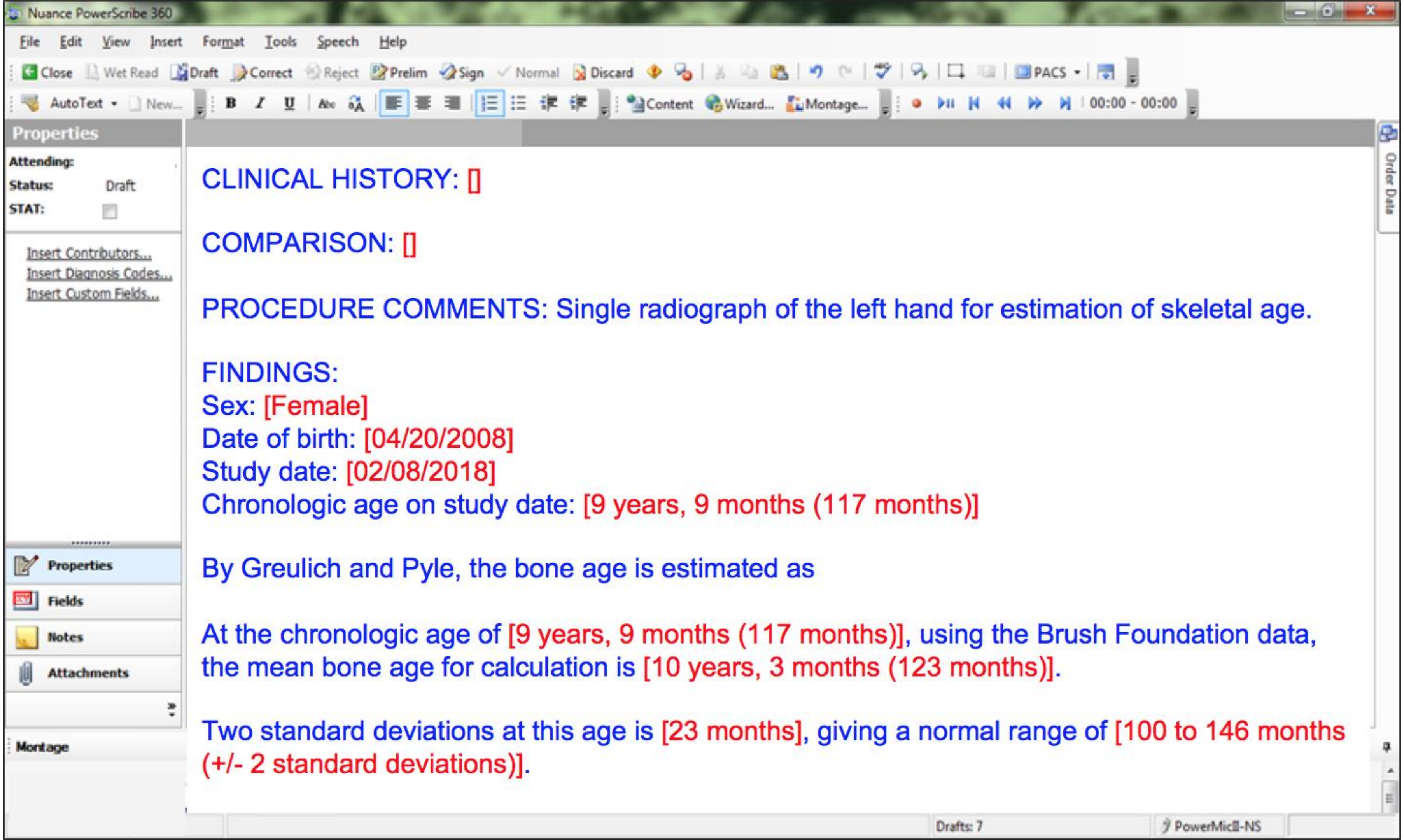
Automation Bias

BONE AGE AI RCT

Prospective study design

- Randomized at the exam level

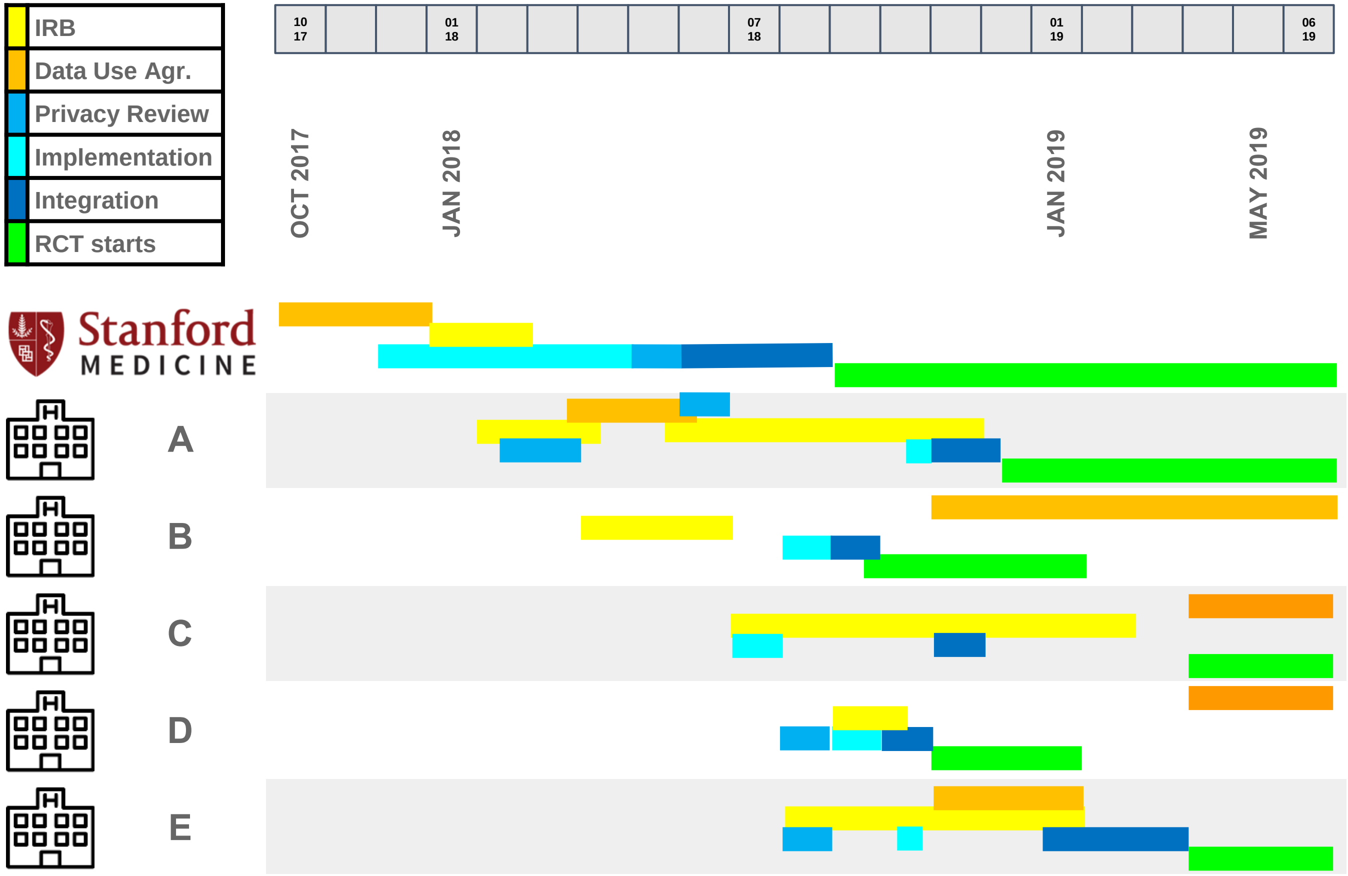



Nuance PowerScribe 360
File Edit View Insert Format Tools Speech Help
Close Wet Read Draft Correct Reject Prelim Sign Normal Discard PACS
AutoText New... Content Wizard... Montage... 00:00 - 00:00
Properties
Attending:
Status: Draft
STAT:
Insert Contributors...
Insert Diagnosis Codes...
Insert Custom Fields...
Properties
Fields
Notes
Attachments
Montage
Order Data
CLINICAL HISTORY: []
COMPARISON: []
PROCEDURE COMMENTS: Single radiograph of the left hand for estimation of skeletal age.
FINDINGS:
Sex: [Female]
Date of birth: [04/20/2008]
Study date: [02/08/2018]
Chronologic age on study date: [9 years, 9 months (117 months)]
By Greulich and Pyle, the bone age is estimated as
At the chronologic age of [9 years, 9 months (117 months)], using the Brush Foundation data, the mean bone age for calculation is [10 years, 3 months (123 months)].
Two standard deviations at this age is [23 months], giving a normal range of [100 to 146 months (+/- 2 standard deviations)].
Drafts: 7
PowerMicII-NS


Allumettes

Abbreviated Timeline of Implementing BA Model at Stanford Children's

10/16 - **Submitted DRA** for review
11/29 - Conference call with **DRA committee** (Lily from ISO, Annie from PO)
12/1 - Meeting with Dr. Halabi in OU; asked for intro to LPCH IS team
12/6 - Meeting with Marvin for **DICOM-SR**
12/8 - Follow-up meeting for DICOM-SR; Requested **firewall change**
12/22 - **DRA approved**
1/3 - **Firewall change approved**
1/9 - **IRB submitted**
1/29 - Modlink can receive my DICOM-SR messages, but cannot interpret them
2/23 - IRB approved
3/5 - Configured LPCH DICOM router to route new studies to the machine learning model
3/28 - Configured Modlink to receive DICOM-SR and tested in test environment; but we need to wait for new Nuance key (at this point, all technical integration work on our end is complete)
4/11 - Received Nuance key; required another firewall change for this key
4/26 - **Firewall change approved**
4/27 - **Change control and additional LPCH security review** for the first time
5/8 - **Security review form** submitted



IRB
Data Use Agr.
Privacy Review
Implementation
Integration
RCT starts
10 17
01 18
07 18
01 19
06 19
OCT 2017
JAN 2018
JAN 2019
MAY 2019
Stanford MEDICINE
A
B
C
D
E

Multi-Institutional Trial

Stanford Children's Health

Children's Hospital of Philadelphia

KU Medical Center
The University of Kansas

Egleston Children's Hospital
at Emory University

NYU Langone Health

Boston Children's Hospital

Yale-New Haven Children's Hospital

MedStar Georgetown University Hospital

Cincinnati Children's

Radiology

ORIGINAL RESEARCH • PEDIATRIC IMAGING

Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment: A Prospective Multicenter Randomized Controlled Trial

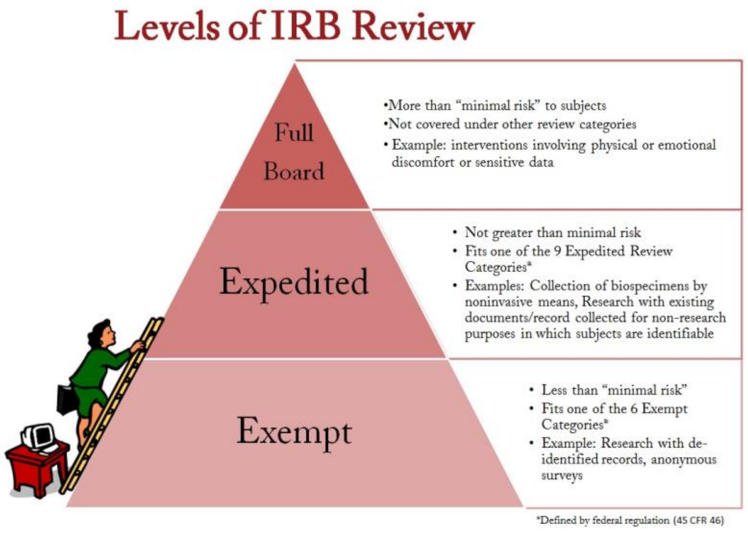
David K. Eng, MS • Nishith B. Khandwala, MS • Jin Long, PhD • Nancy R. Fefferman, MD • Shailee V. Lala, MD • Naomi A. Strubel, MD • Sarah S. Milla, MD • Ross W. Filice, MD • Susan E. Sharp, MD • Alexander J. Towbin, MD • Michael L. Francavilla, MD • Summer L. Kaplan, MD • Kirsten Ecklund, MD • Sanjay P. Prabhu, MD • Brian J. Dillon, MD • Brian M. Everist, MD • Christopher G. Anton, MD • Mark E. Bittman, MD • Rebecca Dennis, DO • David B. Larson, MD, MBA • Jayne M. Seekins, DO • Cicero T. Silva, MD • Arash R. Zandieh, MD • Curtis P. Langlotz, MD, PhD , • Matthew P. Lungren, MD, MPH • Safwan S. Halabi, MD

From the Department of Computer Science, Stanford University, 300 N Pasteur Dr, Stanford, CA 94305 (D.K.E., N.B.K.); Departments of Pediatrics (J.L.) and Radiology (D.B.L., J.M.S., C.P.L., M.P.L., S.S.H.), Stanford University School of Medicine, Stanford, Calif; Department of Radiology, New York University School of Medicine, New York, NY (N.R.F., S.V.L., N.A.S., M.E.B.); Department of Radiology, Emory School of Medicine and Children's Healthcare of Atlanta, Atlanta, Ga (S.S.M.); Department of Radiology, MedStar Health and Georgetown University School of Medicine, Washington, DC (R.W.F., A.R.Z.); Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio (S.E.S., A.J.T., C.G.A.); Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, Pa (M.L.F., S.L.K., R.D.); Department of Radiology, Harvard Medical School and Boston Children's Hospital, Boston, Mass (K.E., S.P.P.); Department of Radiology, Yale School of Medicine, New Haven, Conn (B.J.D., C.T.S.); and Department of Radiology, Kansas University School of Medicine, Kansas City, Kan (B.M.E.). Received January 14, 2021; revision requested March 16; revision received June 24; accepted July 22. **Address correspondence to** D.K.E. (e-mail: *dkeng@stanford.edu*).

IRB

- Average duration to approval: 5 months

- Common Problems:
  - No central IRB; had to be approved at every institution
  - Patient consent



Data Use Agreements (DUA)

- Average duration: 4 months
- Common problems:
  - Legal departments

Security and Privacy Reviews

- Average duration: 1 month

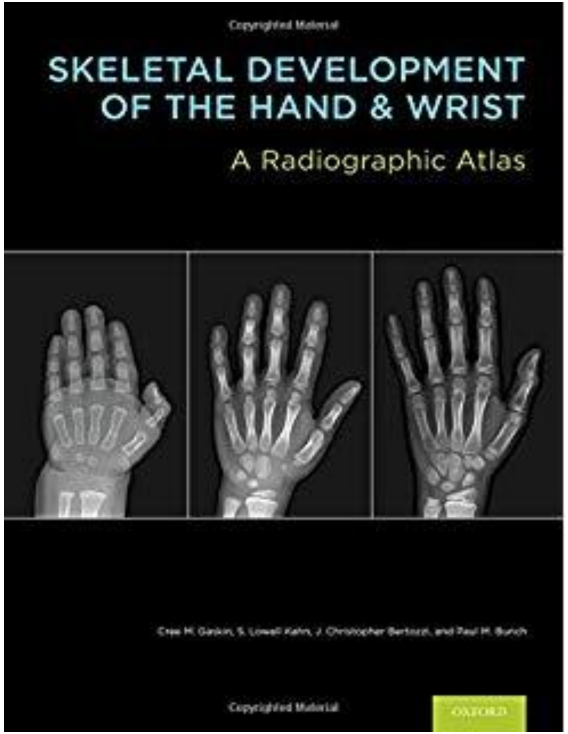
- Common Problems:
  - Latency among information security and privacy offices
  - Lack of clarity in the process

Clinical Integration

- Average duration: 1 month
- Common Problems:
  - Server provisioning
  - Institution-specific interface with speech recognition
  - Institution-specific data preprocessing

Customization Per Site

- Report format varies per site
  - Brush Foundation?
  - Interpolate between bone ages?
  - Greulich & Pyle?



Challenging Clinical Scenarios

- What BA reference should we use?
  - G&P
  - Snell
  - Tanner-Whitehouse
- Does BA model account for brachymetacarpia, dysplasia, malnutrition?
- Does BA model take into account demographics, clinical history, referring clinician practice?

Tracking Performance

- Manual extraction of bone age interpretation from reports

Establishing Gold Standard

- Interpretation by a panel of 4 radiologists for every exam

Overall progress:
86.3% of real's completed (7469/8657)
91.3% of fake's completed (210/230)

Progress per labeler:
alexandertowbin: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
arashzandieh: 100.0% of real's completed (599/599) | 100.0% of fake's completed (10/10)
briandillon: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
chrisanton: 100.0% of real's completed (596/596) | 100.0% of fake's completed (10/10)
cicerosilva: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
jayneseekins: 100.0% of real's completed (595/595) | 100.0% of fake's completed (10/10)
kirstenecklund: 100.0% of real's completed (299/299) | 100.0% of fake's completed (10/10)
markbittman: 100.0% of real's completed (600/600) | 100.0% of fake's completed (10/10)
mattlungren: 100.0% of real's completed (299/299) | 100.0% of fake's completed (10/10)
michaelfrancavilla: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
naomistrubel: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
rebeccadennis: 100.0% of real's completed (598/598) | 100.0% of fake's completed (10/10)
rossfilice: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
safwanhalabi: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
shaileelala: 100.0% of real's completed (299/299) | 100.0% of fake's completed (10/10)
summerkaplan: 100.0% of real's completed (299/299) | 100.0% of fake's completed (10/10)
susansharp: 100.0% of real's completed (300/300) | 100.0% of fake's completed (10/10)
sarahmilla: 99.7% of real's completed (299/300) | 100.0% of fake's completed (10/10)
davidlarson: 77.3% of real's completed (232/300) | 100.0% of fake's completed (10/10)
nancyfefferman: 67.6% of real's completed (200/296) | 0.0% of fake's completed (0/10)
sanjayprabhu: 33.7% of real's completed (101/300) | 100.0% of fake's completed (10/10)
maceverist: 17.3% of real's completed (52/300) | 100.0% of fake's completed (10/10)
ericariedesel: 0.2% of real's completed (1/577) | 0.0% of fake's completed (0/10)

POST-RCT Reflections

- Integration was not the hardest part; **paperwork** was

- Bone age AI validation is a one-off

POST-RCT

Takeaways

- We need a layer of technical and legal infrastructure across institutions to support prospective validation of AI models at scale
- Standards
  - Data
  - Sharing
  - Implementation
  - Clinical practice

Takeaways

Goals to be accomplished for using AI in daily clinical practice

1. AI solutions should address a <u>significant clinical need</u>
2. Technology must perform <u>at least as well</u> as the existing standard approach
3. <u>Substantial clinical testing</u> must validate the new technology
4. New technology should provide <u>improvements</u> in patient outcomes, patient quality of life, practicality in use, and reduce medical costs
5. <u>COORDINATED APPROACH</u> between multiple stakeholders is needed

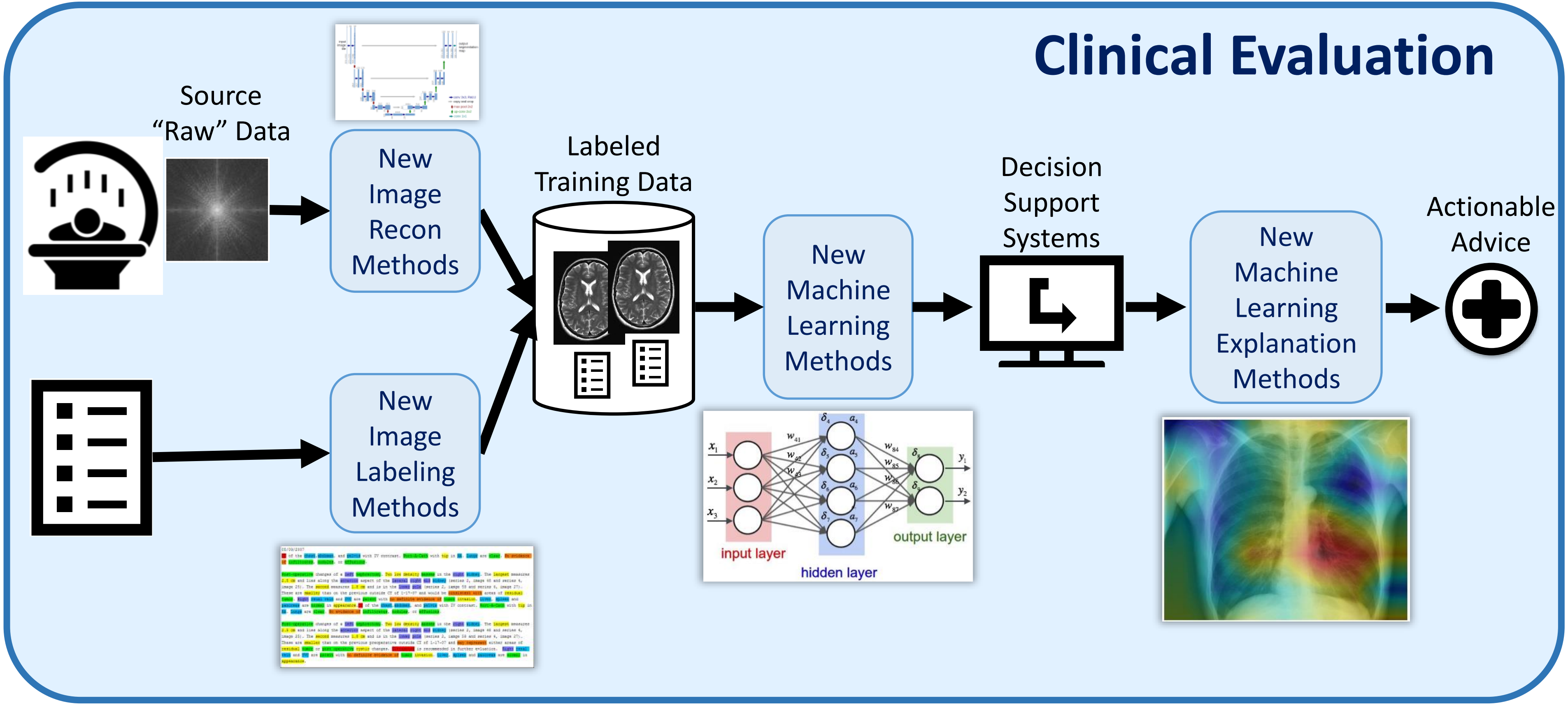
Coordinated Approach

- End users must first define the purpose (clinical use case)
- Developers must translate users' needs to program code
- Managers must coordinate resources and strategies to bring SW in workflow
- Companies must mass distribute the SW product and integrate it with existing infrastructure
- Policy experts and legal teams must ensure there are no legal/ethical barriers

$ Financial Considerations

- Difficult to define a business plan for a narrow AI product that may solve one clinical question on one modality
- May be a pricing disparity between what customers will pay and the costs involved
- Who will pay? Insurance, patient, health system, radiology group, vendor?
- Who is in charge of AI model implementation? Vendor, hospital IS?
- What happens when the model fails or is not fully validated?

Technical Considerations



CT scan icon by Sergey Demushkin from the Noun Project

Building Radiology AI:
The Role of Professional Organizations

- Educate clinical users of AI algorithms
- Develop a robust technical workforce
- Convene collaborations: radiologists, scientists, industry
- Support development of AI use cases
- Assemble publicly-available training data sets
- Advocate for and provide research funding for AI
- Establish standards for AI data and algorithms
- Encourage balanced regulation of AI technology

RSNA®
ACR®
SIIM
ESR

Summary

- AI is a powerful tool with many applications that can help radiology practices today *beyond image interpretation*
- Integrating AI models holds promise for improving radiology practices and patient care
- More research needs to be done regarding the evaluation of AI in a clinical setting, including its impact on workflow and value of services
- No matter how AI is implemented in the workflow, the radiologists will have an important role in ensuring accuracy, safety and quality of the algorithms

Thank you


MICHIGAN
EST. 1817
HAIL TO THE VICTORS
M VS N
SATURDAY | NOON | FOX
CHAMPIONS OF THE WEST!