Challenges of Deploying and Validating an AI Tool into Medical Practice

Safwan S. Halabi MD
Clinical Associate Professor
Department of Radiology

March 19, 2019
Disclosures

Advisor

Bunker Hill
Interfierce (CMO)
DNAFeed

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

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 themselves – 30% of the time

What do radiologists do?
What my family thinks I do
What society thinks I do
What the ER intern thinks I do
What the surgeons think I do
What I think I do
What I actually do
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
What is AI and Why All the Hype?
March of the machines
A SPECIAL REPORT ON ARTIFICIAL INTELLIGENCE

Artificial Intelligence
The promise and the peril
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
Deep Learning

Ability for machines to **autonomously** mimic human thought patterns through artificial neural networks composed of cascading layers of information.
“In the 1970s, an AI system that worked for one patient was worth a masters degree; if it worked for three patients, it was a PhD. Now, it's different.”

--Pete Szolovits, #Peds2040, Jan 2016
AI v1.0: 1950s-1980s

Symbolic Systems

Input → Rule-based systems → Output

Benign
Malignant

AI v2.0: 1980s-2010s

Machine Learning

Input → Feature extraction → Output

Benign
Malignant

AI v3.0: 2010-present

Neural Networks and Deep Learning

Input → Feature extraction → Classification → Output

Benign
Malignant

Not Cancer
Cancer

Neural Networks and Deep Learning
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
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?
Challenge #2: Annotation
Imaging Annotation Value

- Ground Truth
- Prospective Annotation
- Retrospective Annotation
Machine Learning

- Supervised Machine Learning
  - Classification
  - Regression
- Unsupervised Machine Learning
  - Clustering
  - Association
**Classification**
Used when the output is categorical like 'YES' or 'NO'

**Algorithms used**
- Decision Tree
- Naive Bayes
- Random Forest
- Logistic regression
- KNN

**Regression**
Used when a value needs to be predicted like the ‘stock prices’

**Algorithms used**
- Linear Regression

**Clustering**
Used when the data needs to be organized to find patterns in the case of ‘product recommendation’
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
Logistic Regression

Chest radiographs labeled for presence of pneumonia

If greater the 50% of labels or labelers consider image contains pneumonia, then model considers that image positive for pneumonia
Knee MRI Classifier

- Dataset:
  - 1400 knee MRI
  - 3 series
- Labels:
  1. normal/abnormal
  2. ACL tear
  3. Meniscus tear
<table>
<thead>
<tr>
<th>Label</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal</td>
<td>.94</td>
</tr>
<tr>
<td>ACL Tear</td>
<td>.97</td>
</tr>
<tr>
<td>Meniscal Tear</td>
<td>.85</td>
</tr>
</tbody>
</table>
How did we collect MURA?

MURA is a dataset of musculoskeletal radiographs consisting of 14,863 studies from 12,173 patients, with a total of 40,561 multi-view radiographic images. Each image belongs to one of seven standard upper extremity radiographic study types: elbow, finger, forearm, hand, humerus, shoulder, and wrist. Each study was manually labeled as normal or abnormal by board-certified radiologists from the Stanford Hospital at the time of clinical radiographic interpretation in the diagnostic radiology environment between 2001 and 2012.

Test Set Collection

To evaluate models and get a robust estimate of radiologist performance, we collected additional labels from six board-certified Stanford radiologists on the test set, consisting of 207 musculoskeletal studies. The radiologists individually retrospectively reviewed and labeled each study in the test set as a DICOM file as normal or abnormal in the clinical reading room environment using the PACS system. The radiologists have 8.83 years of experience on average ranging from 2 to 25 years. We randomly chose 3 of these radiologists to create a gold standard, defined as the majority vote of labels of the radiologists.

https://stanfordmlgroup.github.io/competitions/mura/
Prospective Labels

1.5M exams labeled prospectively
@ Stanford Radiology

MURA
40k prospectively labeled MSK X-rays
released in 2018 for data challenge
How does our baseline do?

We evaluated our baseline on the Cohen’s kappa statistic, which expresses the agreement of the model with the gold standard. Baseline performance is comparable to radiologist performance in detecting abnormalities on finger studies and equivalent on wrist studies. However, baseline performance is lower than best radiologist performance in detecting abnormalities on elbow, forearm, hand, humerus, shoulder studies, and overall, indicating that the task is a good challenge for future research.

<table>
<thead>
<tr>
<th></th>
<th>Radiologist 1</th>
<th>Radiologist 2</th>
<th>Radiologist 3</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elbow</td>
<td>0.850 (0.830, 0.871)</td>
<td>0.710 (0.674, 0.745)</td>
<td>0.719 (0.685, 0.752)</td>
<td>0.710 (0.674, 0.745)</td>
</tr>
<tr>
<td>Finger</td>
<td>0.304 (0.249, 0.358)</td>
<td>0.403 (0.339, 0.467)</td>
<td>0.410 (0.358, 0.463)</td>
<td>0.389 (0.332, 0.446)</td>
</tr>
<tr>
<td>Forearm</td>
<td>0.796 (0.772, 0.821)</td>
<td>0.802 (0.779, 0.825)</td>
<td>0.798 (0.774, 0.822)</td>
<td>0.737 (0.707, 0.766)</td>
</tr>
<tr>
<td>Hand</td>
<td>0.661 (0.623, 0.698)</td>
<td>0.927 (0.917, 0.937)</td>
<td>0.789 (0.762, 0.815)</td>
<td>0.851 (0.830, 0.871)</td>
</tr>
<tr>
<td>Humerus</td>
<td>0.867 (0.850, 0.883)</td>
<td>0.733 (0.703, 0.764)</td>
<td>0.933 (0.925, 0.942)</td>
<td>0.600 (0.558, 0.642)</td>
</tr>
<tr>
<td>Shoulder</td>
<td>0.864 (0.847, 0.881)</td>
<td>0.791 (0.765, 0.816)</td>
<td>0.864 (0.847, 0.881)</td>
<td>0.729 (0.697, 0.760)</td>
</tr>
<tr>
<td>Wrist</td>
<td>0.791 (0.766, 0.817)</td>
<td>0.931 (0.922, 0.940)</td>
<td>0.931 (0.922, 0.940)</td>
<td>0.931 (0.922, 0.940)</td>
</tr>
<tr>
<td>Overall</td>
<td>0.731 (0.726, 0.735)</td>
<td>0.763 (0.759, 0.767)</td>
<td>0.778 (0.774, 0.782)</td>
<td>0.705 (0.700, 0.710)</td>
</tr>
</tbody>
</table>

https://stanfordmlgroup.github.io/competitions/mura/
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
Machine learning security: These are not stop signs?

Single Pixel Attacks

FDA News Release

FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures

For Immediate Release

May 24, 2018

Summary

FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures
Manufacturer Imagen Technologies of New York City submitted to the FDA a study of 1000 radiographic images that evaluated the software’s independent performance in detecting wrist fractures (OsteoDetect).

Study assessed how accurately the software indicated the location of fractures compared with reviews from 3 board-certified orthopedic hand surgeons.

Also submitted a retrospective study in which 24 clinicians reviewed 200 patient cases.
FDA said both studies showed that sensitivity, specificity, and positive and negative predictive values in detecting wrist fractures improved when clinicians used the software.

- Approved through the FDA’s De Novo regulatory pathway for novel low- to moderate-risk devices.
Imagen OsteoDetect is a type of computer-aided detection and diagnostic software that uses machine learning techniques to identify signs of distal radius fracture during reviews of posterior-anterior and medial-lateral x-ray images of the wrist. Software marks the location of a fracture on the image to aid clinicians with their diagnoses.
Clinicians can use the software in a variety of settings, including primary care, emergency departments, urgent care centers, and for specialty care such as orthopedics.

OsteoDetect is an adjunct tool.

Not meant to replace clinicians’ radiograph reviews or clinical judgment.
There's been a lot of talk about radiologists being replaced by machines. So I looked up the peer-reviewed publications of #AI companies @enlitr @ZebraMedVision @baylabsinc @ArterysInc @radlogics @IBMWatsonHealth
Greatest Potential of AI in HC

Making back-end processes more efficient

10 AI Applications That Could Change Health Care

<table>
<thead>
<tr>
<th>APPLICATION</th>
<th>POTENTIAL ANNUAL VALUE BY 2026</th>
<th>KEY DRIVERS FOR ADOPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robot-assisted surgery</td>
<td>$408</td>
<td>Technological advances in robotic solutions for more types of surgery</td>
</tr>
<tr>
<td>Virtual nursing assistants</td>
<td>20</td>
<td>Increasing pressure caused by medical labor shortage</td>
</tr>
<tr>
<td>Administrative workflow</td>
<td>18</td>
<td>Easier integration with existing technology infrastructure</td>
</tr>
<tr>
<td>Fraud detection</td>
<td>17</td>
<td>Need to address increasingly complex service and payment fraud attempts</td>
</tr>
<tr>
<td>Dosage error reduction</td>
<td>16</td>
<td>Prevalence of medical errors, which leads to tangible penalties</td>
</tr>
<tr>
<td>Connected machines</td>
<td>14</td>
<td>Proliferation of connected machines/devices</td>
</tr>
<tr>
<td>Clinical trial participation</td>
<td>13</td>
<td>Patent cliff; plethora of data; outcomes-driven approach</td>
</tr>
<tr>
<td>Preliminary diagnosis</td>
<td>5</td>
<td>Interoperability/data architecture to enhance accuracy</td>
</tr>
<tr>
<td>Automated image diagnosis</td>
<td>3</td>
<td>Storage capacity; greater trust in AI technology</td>
</tr>
<tr>
<td>Cybersecurity</td>
<td>2</td>
<td>Increase in breaches; pressure to protect health data</td>
</tr>
</tbody>
</table>

SOURCE: ACCENTURE

AI in Radiology: Current State

• **Individual** AI software developers are currently working with individual radiologists at single institutions to create AI algorithms that are focused on targeted interpretive needs.

• Developers are using a single institution’s prior imaging data for training and testing the algorithms, and the algorithm output is specifically tailored to that site’s perspective of the clinical workflow.

• Will models be generalizable to widespread clinical practices?

• How will model be integrated into clinical workflows across a variety of practice settings?

Advancing AI Algorithms for Radiology

• “Ensuring that algorithms can be integrated into radiologists’ clinical workflow is of paramount importance because if the AI tool is not readily available to the end users in their workflow, adoption in clinical practice will be less likely to occur.”
  (B. Allen, K. Dreyer)

• Interoperability between all systems is prerequisite

• Radiologists have to chose the best model for implementing AI
  • How to activate AI analysis and for what purpose
  • How to incorporate image analysis results in their reports

M. Walter, Radiology Business, May 07, 2018
B. Allen, JACR, DOI: https://doi.org/10.1016/j.jacr.2018.02.032
Implementing AI in Radiology

- Developers of AI algorithms do not always have a strong medical background or understanding of physician workflow
- Lack of well curated and diverse datasets
- "You have to have validated data sets to train [the algorithms], and so the use cases now are just being driven by data availability, not by cases that people care about. No one cares about bone age" (Paul Chang MD)
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
Implementing AI: 3 Possible scenarios

1. AI on demand
2. Automated image analysis
3. Discrepancy management
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 before final report is made
   • Radiologist is able to ensure accuracy
Scenario 3

3. Discrepancy management

- As in 2. but results are automatically routed to RIS or EHR
- Requires discrepancy management
- 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
Bone Age The Old Way

A Depeursinge et al, Open Medical Informatics Journal 11:2017

Measuring Delayed Growth
Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs

Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Clinical Report</th>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>Reviewer 3</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD Reviewer</td>
<td>0.65</td>
<td>0.55</td>
<td>0.53</td>
<td>0.69</td>
<td>0.61</td>
</tr>
<tr>
<td>Model</td>
<td>0.51</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td><em>P</em> value (paired <em>t</em> test)</td>
<td>&lt;.01</td>
<td>.50</td>
<td>.99</td>
<td>&lt;.01</td>
<td></td>
</tr>
</tbody>
</table>

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

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).
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?
Control group

Diagnose using standard of care

Expt. group

Diagnose using the help of the AI
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)].
Validation Design
Scenarios

- Scenario 1: Popup window with recommendation and prediction?
- Scenario 2: Prepopulate report?
- Scenario 3: Automatically publish report?
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
Clinical Scenarios

• Quick question since you do a lot of bone age stuff. Patient JG 13y8m genetic female, transitioning to male and on hormone therapy. What is current practice in reporting in these cases? We are just going to report bone age for both genders. Thoughts?
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?
Multi-Institutional Trial

Stanford Children's Health
240

Egleston Children's Hospital
300

NYU Langone Health
450

Boston Children's Hospital

Yale-New Haven Children's Hospital

MedStar Georgetown University Hospital
80

Cincinnati Children's
Key Recommendations

Goals to be accomplished for using AI in daily clinical practice

1. AI solutions should address a significant clinical need
2. Technology must perform at least as well as the existing standard approach
3. Substantial clinical testing must validate the new technology
4. New technology should provide improvements in patient outcomes, patient quality of life, practicality in use, and reduce medical costs
5. COORDINATED APPROACH 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
Who are the Stakeholders?

HC Community
• Radiologists and residents/trainees
• Referring physicians and patients
• Medical professional societies
• Hospital systems, IT departments
• Academics and medical scientists

SW Community
• IT professionals, SW developers
• Health information technology (HIT) industry
• Academic IT professionals: engineers, computer scientists
Other Stakeholders

• Governments and insurance companies
  • Financing, reimbursement
  • Different payment models (public, hybrid)
  • Variable strategies for fostering AI software in general and for HC
• Regulatory agencies (FDA, CE)
• Patients
$ 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?
Clinical Evaluation

Source “Raw” Data

New Image Recon Methods

New Labeled Training Data

New Image Labeling Methods

Labeled Training Data

Decision Support Systems

New Machine Learning Methods

New Machine Learning Explanation Methods

Actionable Advice

CT scan icon by Sergey Demushkin from the Noun Project
AI and the Radiologist

- How does the AI algorithm influence the performance of the radiologist?
- Does Radiologist + AI outperform just the Radiologist?
- What is considered the “ground truth”?
- How will the AI model be displayed?
- Will the AI model learn over time?
I'm being sued for a missed diagnosis! What do I do now?!

Sorry, buddy, comes with the job.

When handing radiology over to artificial intelligence sounds appealing.
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
Take Home Messages

• 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
Will #AI replace radiologists? The answer is NO. But rads who use #AI will replace rads who don't  
@RSNAInformatics @SIIM_Tweets
David Eng • 1st
Student at Stanford University
Stanford, California

Nishith Khandwala • 1st
Machine Learning Researcher at Stanford University
Stanford, California
Questions?

Safwan.Halabi@Stanford.edu

@SafwanHalabi