서울아산병원
Asan Medical Center

Transparency of AI from User's Perspective

AIRIS 2024, Feb. 27, 2024, Seoul

Seong Ho Park, MD, PhD
Professor
Asan Medical Center, University of Ulsan

Disclosure

- No conflicts of interest with any AI products or vendors included in the presentation.

Contents

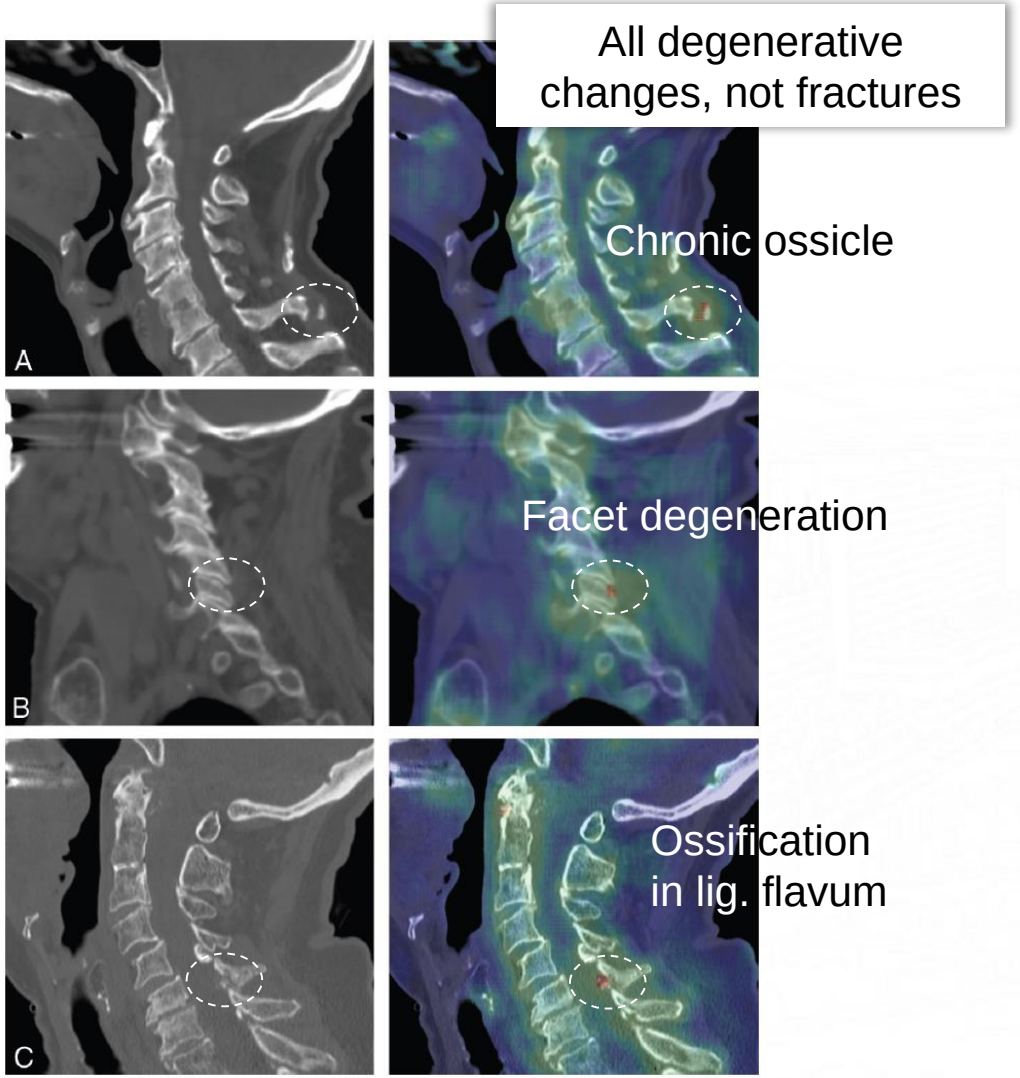
Transparency from the user’s perspective including

1) Model performance and data
2) Trustworthiness of AI predictions
3) Responsible human supervision in the use of AI

To elucidate the relevance of these and suggest what regulatory bodies should do further to enhance transparency in these areas

1. Transparency regarding model performance and data

AI for detection of cervical spine fracture on CT

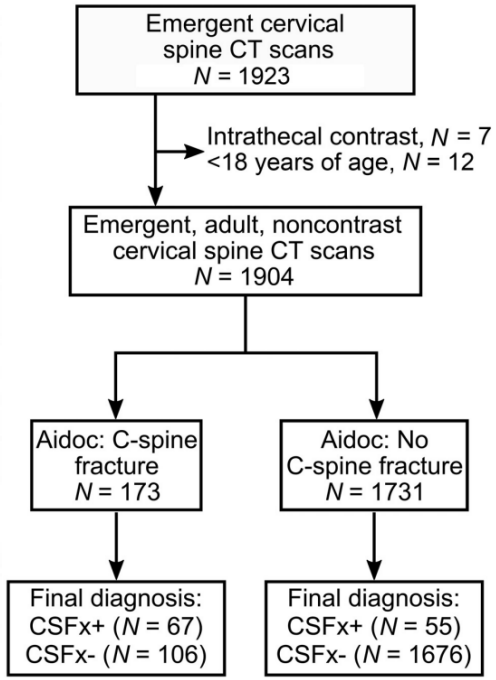


U.S. FOOD & DRUG ADMINISTRATION

Sensitivity and specificity exceeded the 80% performance goal. Specifically, sensitivity was 91.7% (95% CI: 82.7%, 96.9%) and specificity was 88.6% (95% CI: 81.2%, 93.8%).

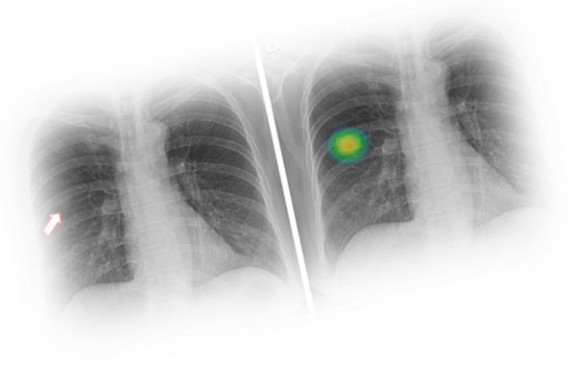
sensitivity was 91.7%

specificity was 88.6%



Sensitivity: 54.9%
Specificity: 94.1%
PPV: 38.7%

Voter et al. *AJNR Am J Neuroradiol*. 2021;42(8):1550-1556



Commercial AI for CXR

Nam et al. AI Improves Nodule Detection on Chest Radiographs in a **Health Screening Population**: A Randomized Controlled Trial. *Radiology*. 2023 Apr;307(2):e221894.

- Seoul National University
- n=10476
- *"In health checkup participants, artificial intelligence–based software improved the detection of actionable lung nodules on chest radiographs."*

Kim et al. Multicentre external validation of a commercial artificial intelligence software to analyse chest radiographs in **health screening environments** with low disease prevalence. *Eur Radiol*. 2023 May;33(5):3501-3509.

- Korea University
- n=3047
- AUROC: 0.648
- Sensitivity: 35.3%
- Specificity: 94.2%
- *"The mean reading time was 2.96–10.27 s longer with AI assistance."*

Limited generalizability of AI in healthcare

- The myth of generalisability in clinical research and machine learning in health care.[1]
- Clinical prediction models are never truly validated due to expected heterogeneity in model performance between locations and settings, and over time.[2]
- The purpose of external testing of an AI algorithm is not to prove its universal generalizability.[3]

1. Futoma et al. *Lancet Digit Health* 2020;2(9):e489-e492
2. Van Calster et al. *BMC Med* 2023;21(1):70
3. Park et al. *Radiology* 2023;306(1):20-31

- Regulatory approval (such as USFDA or Korea MFDS) of an AI as a medical device does not necessarily mean it’s ready for use in everyone’s clinical practice.
- How can a user know more transparently how an AI would work in the user’s practice?

Multi-site external evaluation for regulatory approval

- For 130 AI devices approved by the USFDA (Jan. 2015–Dec. 2020)[1]
  - No multi-site assessment in 93
  - Two-site assessment in 8
- An AI model that exhibits good performance in populations at multiple sites may not perform well at the next site, or vice versa.[2]

1. Wu et al. *Nat Med* 2021;27(4):582–584.
2. Park et al. *Radiology* 2023;308(3):e230288.

Perhaps, greater transparency regarding data is helpful and more effective.

- Sufficient on-site testing before adoption of AI in the user's practice is ideal but not always achievable.
- ***Data transparency***:

  If the user knows whether training and testing data are similar or dissimilar to the data in the user's practice where the AI is intended to be used…
- Further efforts to improve data transparency for end users

Suggesting "model facts" for AI end users in addition to device approval summary, similar to package inserts for drugs

- data
- indications
- proper usage

Model Facts **Model name:** Deep Sepsis **Locale:** Duke University Hospital

Approval Date: 09/22/2019 **Last Update:** 01/13/2020 **Version:** 1.0

Summary

This model uses EHR input data collected from a patient's current inpatient encounter to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.

Mechanism

- **Outcome** ..... sepsis within the next 4 hours, see outcome definition in "Other Information"
- **Output** ..... 0% - 100% probability of sepsis occurring in the next 4 hours
- **Target population** ..... all adult patients >18 y.o. presenting to DUH ED
- **Time of prediction** ..... every hour of a patient's encounter
- **Input data source** ..... electronic health record (EHR)
- **Input data type** ..... demographics, analytes, vitals, medication administrations
- **Training data location and time-period** ..... DUH, diagnostic cohort, 10/2014 – 12/2015
- **Model type** ..... Recurrent Neural Network

Validation and performance

	Prevalence	AUC	PPV @ Sensitivity of 60%	Sensitivity @ PPV of 20%	Cohort Type	Cohort URL / DOI
Local Retrospective	18.9%	0.88	0.14	0.50	Diagnostic	arxiv.org/abs/1708.05894
Local Temporal	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182
Local Prospective	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**
External	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**	**TBD**
Target Population	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182

Uses and directions

- **Benefits:** Early identification and prompt treatment of sepsis can improve patient morbidity and mortality.
- **Target population and use case:** Every hour, data is pulled from the EHR to calculate risk of sepsis for every patient at the DUH ED. A rapid response team nurse reviews every high-risk patient with a physician in the ED to confirm whether or not to initiate treatment for sepsis.
- **General use:** This model is intended to be used to by clinicians to identify patients for further assessment for sepsis. The model is not a diagnostic for sepsis and is not meant to guide or drive clinical care. This model is intended to complement other pieces of patient information related to sepsis as well as a physical evaluation to determine the need for sepsis treatment.
- **Appropriate decision support:** The model identifies patient X as at a high risk of sepsis. A rapid response team nurse discusses the patient with the ED physician caring for the patient and they agree the patient does not require treatment for sepsis.
- **Before using this model:** Test the model retrospectively and prospectively on a diagnostic cohort that reflects the target population that the model will be used upon to confirm validity of the model within a local setting.
- **Safety and efficacy evaluation:** Analysis of data from clinical trial (NCT03655626) is underway. Preliminary data shows rapid response team, nurse-driven workflow was effective at improving sepsis treatment bundle compliance.

Warnings

- **Risks:** Even if used appropriately, clinicians using this model can misdiagnose sepsis. Delays in a sepsis diagnosis can lead to morbidity and mortality. Patients who are incorrectly treated for sepsis can be exposed to risks associated with unnecessary antibiotics and intravenous fluids.
- **Inappropriate Settings:** This model was not trained or evaluated on patients receiving care in the ICU. Do not use this model in the ICU setting without further evaluation. This model was trained to identify the first episode of sepsis during an inpatient encounter. Do not use this model after an initial sepsis episode without further evaluation.
- **Clinical Rationale:** The model is not interpretable and does not provide rationale for high risk scores. Clinical end users are expected to place model output in context with other clinical information to make final determination of diagnosis.
- **Inappropriate decision support:** This model may not be accurate outside of the target population, primarily adults in the non-ICU setting. This model is not a diagnostic and is not designed to guide clinical diagnosis and treatment for sepsis.
- **Generalizability:** This model was primarily evaluated within the local setting of Duke University Hospital. Do not use this model in an external setting without further evaluation.
- **Discontinue use if:** Clinical staff raise concerns about utility of the model for the indicated use case or large, systematic changes occur at the data level that necessitates re-training of the model.

Other information:

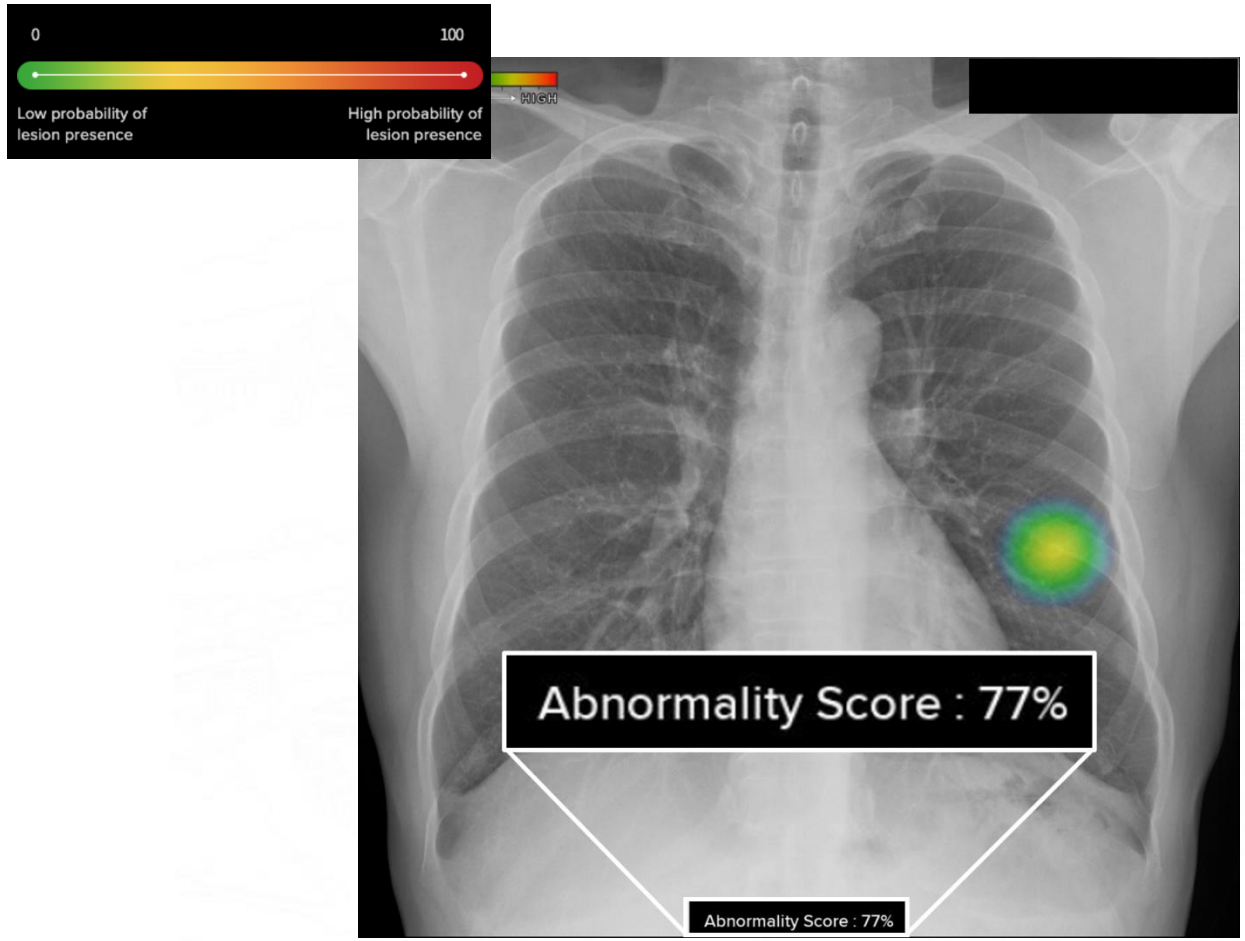
- **Outcome Definition:** https://doi.org/10.1101/648907
- **Related model:** http://doi.org/10.1001/jama.2016.0288
- **Model development & validation:** arxiv.org/abs/1708.05894
- **Model implementation:** jmir.org/preprint/15182
- **Clinical trial:** clinicaltrials.gov/ct2/show/NCT03655626
- **Clinical impact evaluation:** TBD
- **For inquiries and additional information:** please email mark.sendak@duke.edu

Sendak et al. *NPJ Digit Med*. 2020 Mar 23;3:41.

2. Transparency regarding trustworthiness of AI predictions

How can users determine the trustworthiness of an AI prediction?

https://www.lunit.io/en/products/cxr



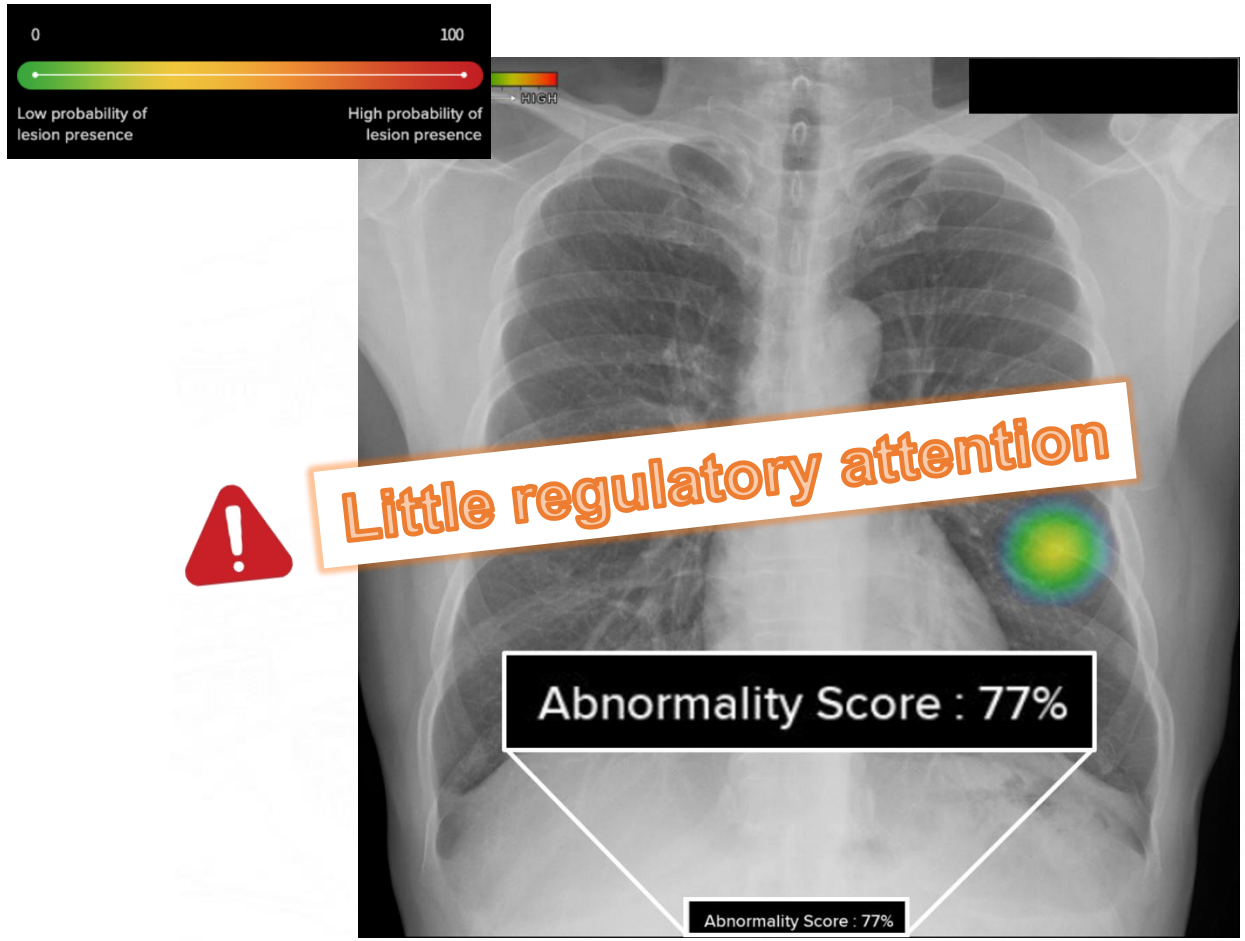
Abnormality/probability score…?

- 77% probability of the target disease?
- 77% certainty that the disease is present?
- Can we trust the AI result more when the score is higher?

Answer: Not really

How can users determine the trustworthiness of an AI prediction?

https://www.lunit.io/en/products/cxr



Abnormality/probability score[1]

- raw AI output before applying threshold
- not or cannot be calibrated[1,2]
- not considering pretest probability[1]
- not a certainty[3]
  - “90% probability of rain, but I am not certain”
  - “20% probability of rain, and I am certain”

1. https://doi.org/10.3348/kjr.2024.0144
2. Van Calster et al. *BMC Med* 2023;21(1):70
3. Faghani et al. *Radiology* 2023;308(2):e222217

How can users determine the trustworthiness of an AI prediction?

Uncertainty quantification (measure of uncertainty)[1]

- Currently at research stage
- An area to which regulatory bodies may need to give more attention in the future.
- Calibration (for probability) alone does not measure uncertainty.
- In addition to reporting an outcome probability, disclosing the prediction uncertainty is essential for user transparency regarding trustworthiness of AI prediction.

1. Faghani et al. *Radiology* 2023;308(2):e222217

3. Transparency regarding responsible human supervision in the use of AI

Proper human supervision is critical.

- For AI to provide real benefits, its use should avoid both automation bias (AI alone) and AI being noninformative redundancy/formality (human alone).
- A synergistic integration of human and AI strengths can be promoted by enhanced transparency regarding responsible human supervision.
- A separate keeping of AI predictions (with a digital watermark, especially for generative AI) and the final clinical decision in the form of a signed medical note or report can improve transparency regarding responsible human supervision.
- An area relevant to both device approval and post-approval stages.
- At the device/regulatory approval level, is there anything that can be done to enhance transparency?

Thank you for your attention.