Is my model applicable, valid, interpretable, and useful?

Ryeyan Taseen, MD, MSc, FRCPC

Background

- MD at Université de Sherbrooke
- Residency in Internal Medicine/Respirology
- MSc in Computer Science/Clinician Scientist Fellowship
- Masters thesis on the development, validation and the evaluation of the utility of a ML model for predicting mortality among hospitalized patients

Objective

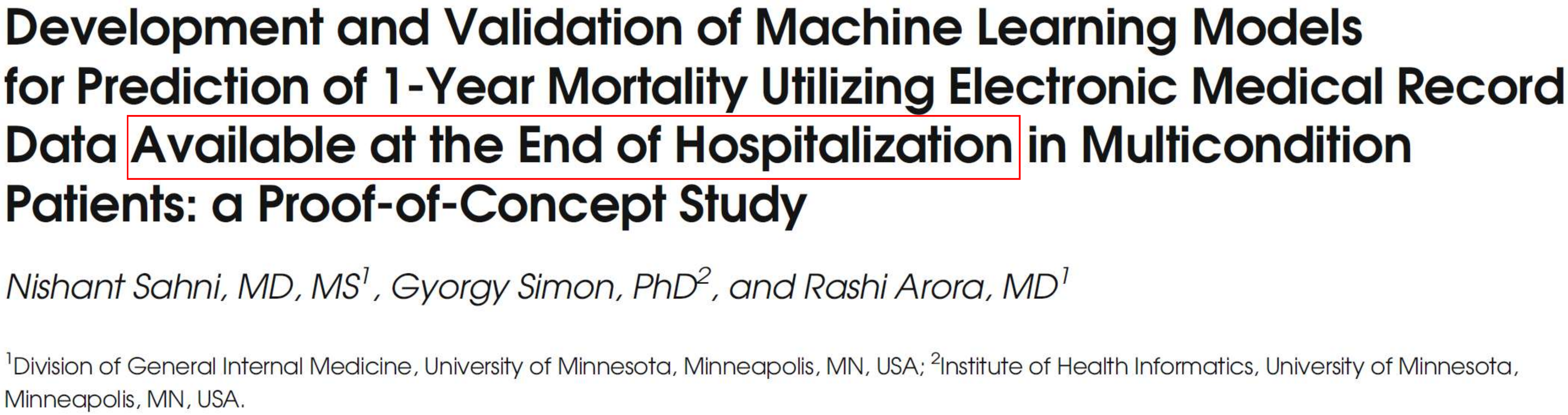
- Provide an overview of the clinical requirements of AI models in health care
  - How will a clinical stakeholder view my model?
- What are the risks to adoption?
  - Model is not applicable
  - Model is not valid enough
  - Model is not interpretable enough
  - Model is not useful enough

Necessary properties of AI for health care

- Applicability
- Validity
- Interpretability
- Utility

AI models must be applicable

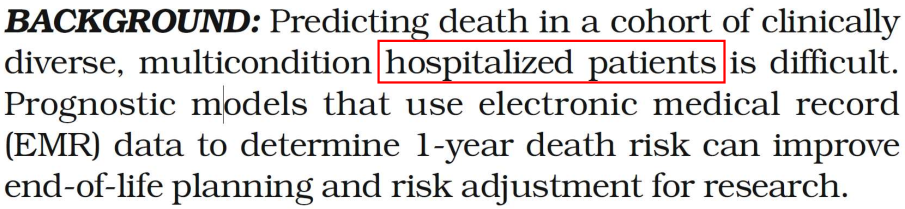
- A model must be able to output predictions at the time and place of the decisions that are intended to be supported



Development and Validation of Machine Learning Models for Prediction of 1-Year Mortality Utilizing Electronic Medical Record Data Available at the End of Hospitalization in Multicondition Patients: a Proof-of-Concept Study

Nishant Sahni, MD, MS[1], *Gyorgy Simon, PhD*[2], *and Rashi Arora, MD*[1]

[1]Division of General Internal Medicine, University of Minnesota, Minneapolis, MN, USA; [2]Institute of Health Informatics, University of Minnesota, Minneapolis, MN, USA.



BACKGROUND: Predicting death in a cohort of clinically diverse, multicondition hospitalized patients is difficult. Prognostic models that use electronic medical record (EMR) data to determine 1-year death risk can improve end-of-life planning and risk adjustment for research.

Model intended to predict mortality for hospitalized patients
Model requires data that is only available after hospitalization
→Model not applicable for inpatient decision support

AI models must be applicable

- At risk whenever input features include diagnoses, notes, reports or other narrative inputs that might not be available in real-time.
- To mitigate, ensure alignment of
  - Clinical decision support process (e.g., Admission day 1)
  - Data generation process (e.g., creation of relevant reports, resulting of labs)
  - Data access process (e.g., access from real-time repository)
- Retrospective AI model validation should try to reproduce the actual data available at the time of decision support

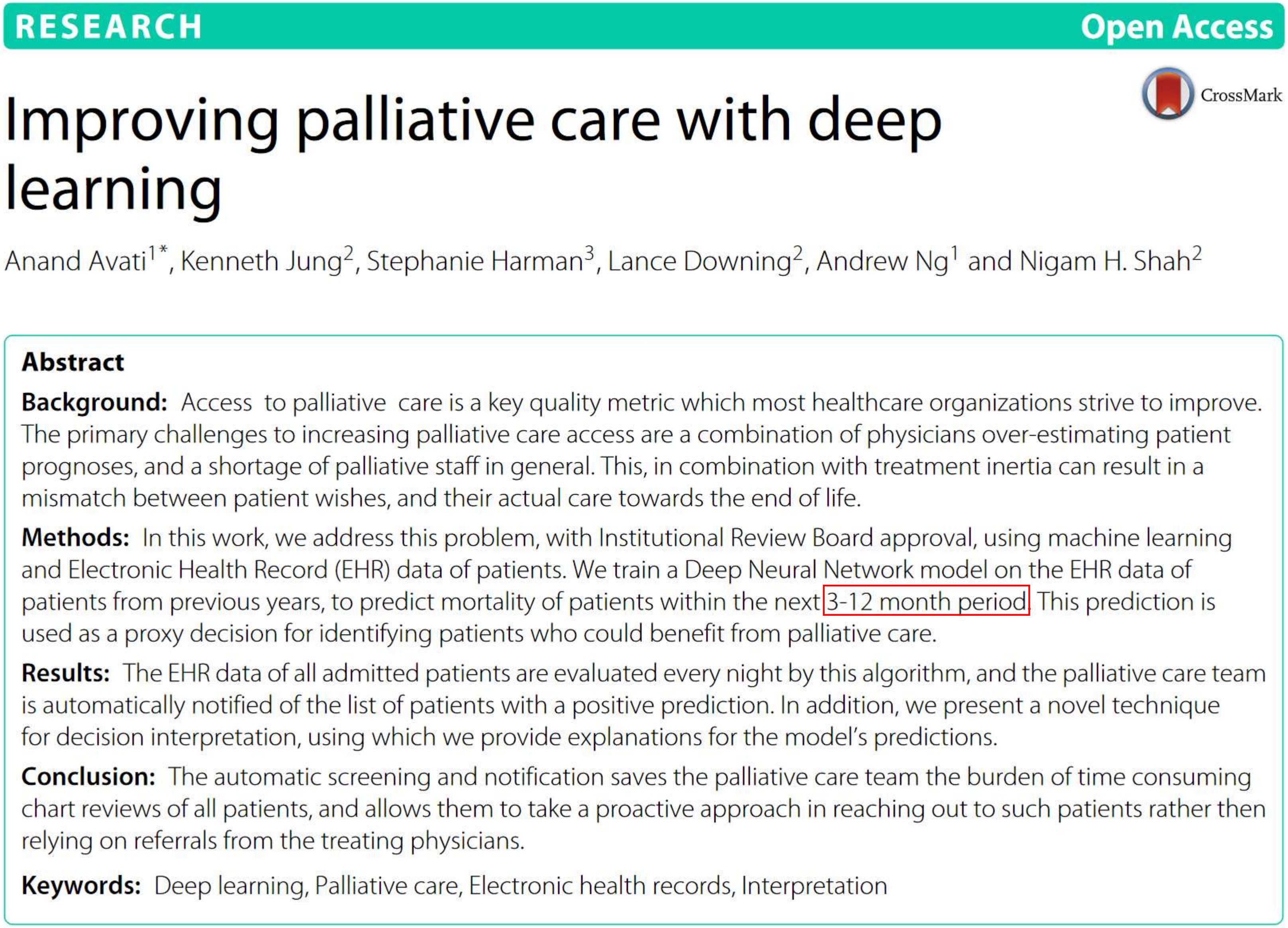
AI models must be valid

- Models need to output good predictions in the intended population
- Validation methods should reflect the clinical context
  - Choice of performance metrics to report/optimize
  - Choice of alternative models to compare with (including expert predictions)
  - Choice of patient groups: training cohort, testing cohort, target population
- Standard PICO framework for appraising research by clinicians:
  - Population: do my patients belong to this group?
  - Intervention (ML model): is this applicable locally?
  - Comparison: Is the intervention compared with the current standard?
  - Outcome: Is the performance metric relevant to practice?

AI models must be valid: localizing validity

- Any report about model validity must reference the population in which validity was tested and the type of validation performed
  - Target population: the model intends to generalize to this group
  - Accessible population: the population from which the validation cohort is sampled
  - Are there major differences between the target and accessible population?

- Mismatch between the included cohort and intended target population can be subtle…



RESEARCH Open Access

Improving palliative care with deep learning

Anand Avati[1*], Kenneth Jung[2], Stephanie Harman[3], Lance Downing[2], Andrew Ng[1] and Nigam H. Shah[2]

Abstract

Background: Access to palliative care is a key quality metric which most healthcare organizations strive to improve. The primary challenges to increasing palliative care access are a combination of physicians over-estimating patient prognoses, and a shortage of palliative staff in general. This, in combination with treatment inertia can result in a mismatch between patient wishes, and their actual care towards the end of life.

Methods: In this work, we address this problem, with Institutional Review Board approval, using machine learning and Electronic Health Record (EHR) data of patients. We train a Deep Neural Network model on the EHR data of patients from previous years, to predict mortality of patients within the next 3-12 month period. This prediction is used as a proxy decision for identifying patients who could benefit from palliative care.

Results: The EHR data of all admitted patients are evaluated every night by this algorithm, and the palliative care team is automatically notified of the list of patients with a positive prediction. In addition, we present a novel technique for decision interpretation, using which we provide explanations for the model's predictions.

Conclusion: The automatic screening and notification saves the palliative care team the burden of time consuming chart reviews of all patients, and allows them to take a proactive approach in reaching out to such patients rather then relying on referrals from the treating physicians.

Keywords: Deep learning, Palliative care, Electronic health records, Interpretation

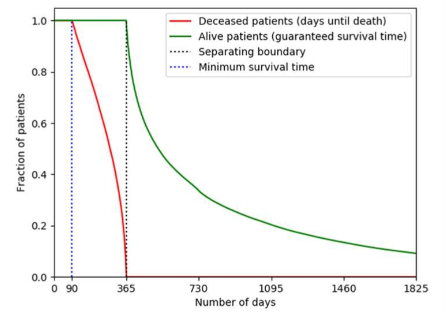


Fig. 1 Right-censoring lengths shown as a survival plot

Patients with mortality < 3 months were excluded from analysis

Not reproducible in real-time, therefore the validation metrics are not representative of real-time use

AI models must be valid: localizing validity

- Type of validation:
  - Internal validity: performance of the model in the train/test/tune cohort
  - External validity: performance of the model in a cohort whose data has not been used for training or tuning.
- The purpose of external validation is to show the model is **generalizable** to the setting it is intended to be deployed in

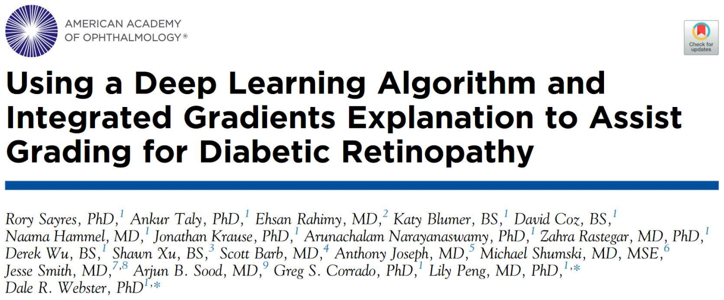
Recommended reading: Futoma et al. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Health*, 2020

AI models must be valid: measuring validity

- Areas under curves and other “overall performance” metrics are useful to compare models with each other, but don’t hold much clinical meaning
  - In actual practice, models operate at a specific threshold or a select few threshold along the AUROC or AUPRC; most of the “area” is irrelevant to patient care.
  - Work with clinical stakeholders to identify a meaningful threshold range for the intended application
  - Report prediction metrics at those thresholds
- Assess performance metrics for all relevant subgroups
  - Very useful for identifying biases

AI models must be valid: relativizing validity

- Model performance should be compared with the performance of validated alternatives
  - E.g., any currently used models for the same task (like clinical scores)
  - E.g., the performance of experts given the same task of classification
- Keep in mind the goal isn’t necessarily to beat the expert
  - In many applications, the goal is to aid clinicians, not replace them.
  - A meaningful comparison group might be AI-assisted expert classification.



AMERICAN ACADEMY
OF OPHTHALMOLOGY®

Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy

Rory Sayres, PhD,[1] Ankur Taly, PhD,[1] Ehsan Rahimy, MD,[2] Katy Blumer, BS,[1] David Coz, BS,[1] Naama Hammel, MD,[1] Jonathan Krause, PhD,[1] Arunachalam Narayanaswamy, PhD,[1] Zahra Rastegar, MD, PhD,[1] Derek Wu, BS,[1] Shawn Xu, BS,[3] Scott Barb, MD,[4] Anthony Joseph, MD,[5] Michael Shumski, MD, MSE,[6] Jesse Smith, MD,[7,8] Arjun B. Sood, MD,[9] Greg S. Corrado, PhD,[1] Lily Peng, MD, PhD,[1,] Dale R. Webster, PhD[1,*]*

AI models must be interpretable

- Many definitions in the literature
- **A practical guide:** being able to verify model desiderata that are not formalized in the learning objective
  - Has the model learned **causal** relationships?
  - Are predictions **transferable** to a dynamic environment?
  - Is model output **informative** enough for decision makers?
  - Are actions based on the predictions **fair (i.e., unbiased relative to vulnerable populations)**?
- If a model is not interpretable, a user can’t verify if it satisfies these requirements for a given application
  - Methods to increase interpretability are methods that help users verify that these requirements are met, and conversely, identify cases where these requirements are not met

Lipton, The Mythos of Model Interpretability, Arxiv, 2016

Interpretability depends on the use case

- Different **applications** have different **requirements**
  - Independent of **model**
- E.g., a model that predicts mortality can be used for
  - Prompting a discussion about code status
  - Prompting an ICU consultation
  - Triaging scarce resources in a pandemic…
- E.g., a model that identifies lung cancer might be used for:
  - Assisting a radiologist
  - Replacing a radiologist
- Determining the importance of interpretability for a given use case:
  - What is at stake if the model is wrong?

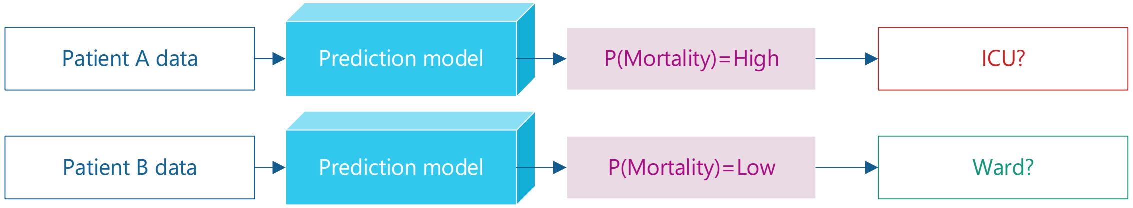
Causality

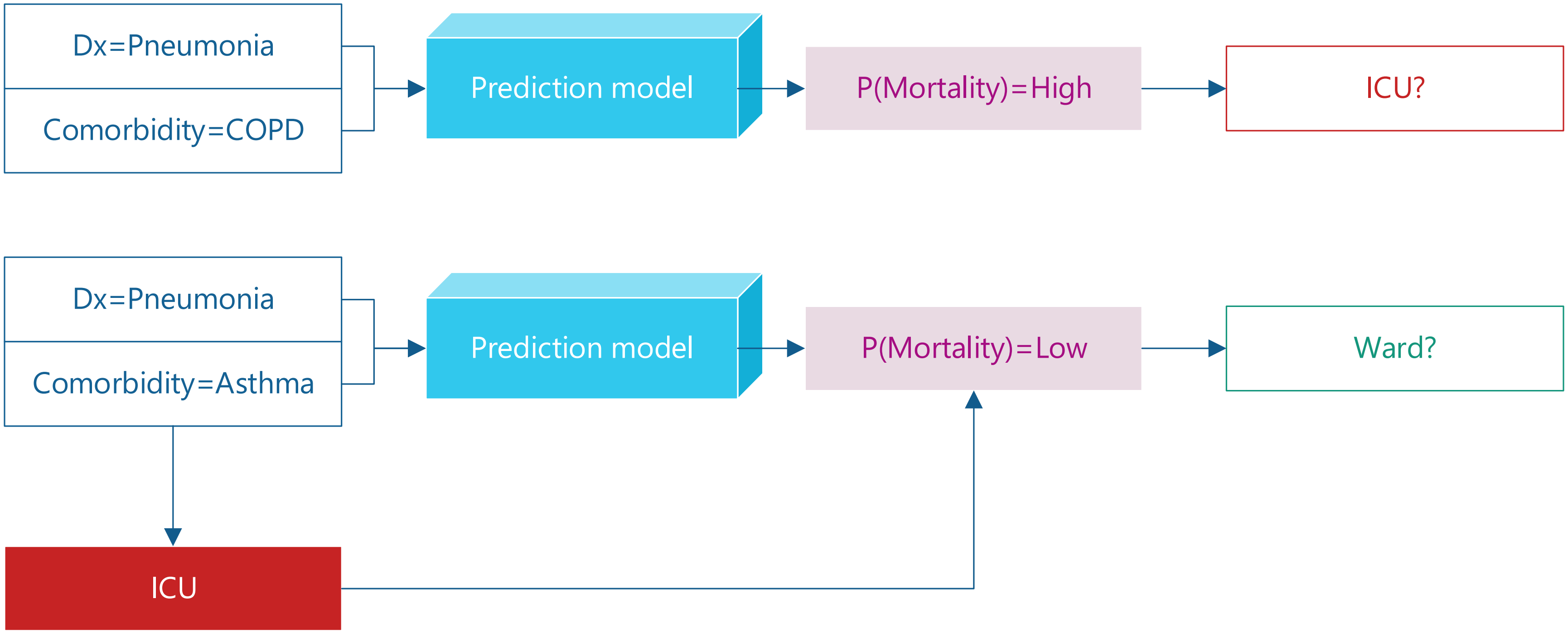
- With traditional prediction models, e.g. using regression, the input features are manually selected and causal relationships are a factor in selection
  - Cholesterol, blood pressure and age for predicting CV risk
- Extreme number of features in more flexible ML methods increase the chance of confounding
  - E.g., pneumonia based on type of machine used for taking X-ray
  - E.g., melanoma based on presence of ruler/skin markings
  - E.g., fracture based on priority marking on x-ray
- Much of the need for interpretability stems from this issue
- The goal of explanations is often to allow experts to detect some confounding
  - I.e., identify undue importance given to non-causal features

Kelly et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 2019
Prosperi et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare, Nature Machine Intelligence, 2019

Transferability

- Consider an AI model to predict mortality for patients with pneumonia in order to prompt ICU admission
- Input (X) = clinical variables (diagnosis, labs, medication)
- Output (Y) = probability of mortality at 30 days
- Excellent predictive performance
- Applied: Patient A has a higher predicted mortality than Patient B. Should Patient A be treated more aggressively than Patient B?





Ahmad et al. Interpretable Machine Learning in Healthcare. IEEE Intelligent Informatics Bulletin. 2018

Transferability

- Patient B might have a lower observed risk of mortality, but this might be because of a condition that has a high risk of mortality when untreated
  - E.g., asthma exacerbation, diabetic ketoacidosis
- I.e., the probability of mortality we trained the model to predict assumes that care would proceed as usual.
- Deviating from routine care would invalidate the model (and harm patients)

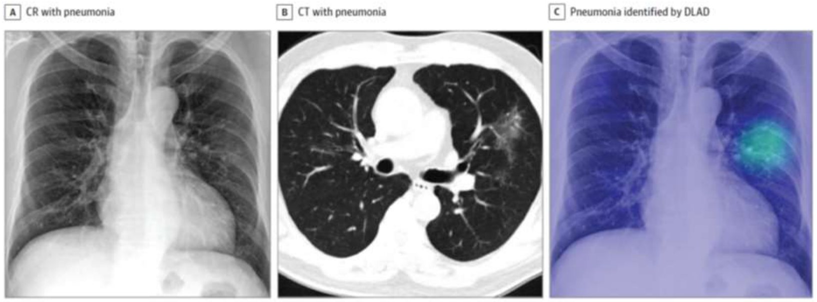
Informativeness

- Consider an AI model to predict pneumonia in order to start earlier management
- Input (X) = Chest X-ray
- Output (Y) = Probability of pneumonia
- Labels obtained from expert readings
- Excellent prediction accuracy
- Good enough for a clinician?

37% chance of pneumonia

Informativeness

Figure 3. Representative Case From the Observer Performance Test (Pneumonia)



A, The chest radiograph (CR) shows subtle patchy increased opacity at the left middle lung field, which was initially missed by all 15 observers. B, The corresponding computed tomographic (CT) image shows patchy ground glass opacity at the left upper lobe. C, The deep learning-based automatic detection algorithm (DLAD) correctly localized the lesion (probability score, 0.371). Seven observers correctly detected the lesion after checking the result.

- Management is also impacted by location of pneumonia, associated complications on imaging, comparison with prior images, differential diagnosis…
- Limited output from the model is not “interpretable” if the user can’t meaningfully use the information

[Hwang et al. Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Open, 2019]

Fairness

- Consider an AI model to predict the need for palliative care consultation
- Input (X) = clinical variables
- Output (Y) = probability of palliative care consult
- I.e., historical consults are used to predict future consults
- Excellent prediction accuracy

- What is the risk of using the model to automatically flag patients for palliative care team review?

Murphree et al. Improving the delivery of palliative care through predictive modeling and healthcare informatics, JAMIA, 2021

Fairness

- What if social inequities resulted in less access to palliative care for a given group?
  - Ethnicity, language, gender
- The model learns that patients belonging to that group are less likely to have palliative care
- When applied, patients belonging to that group are less likely to be flagged and less likely to benefit from future palliative care
- Retraining the model in the future would further propagate the bias

Porter et al. Power and perils of prediction in palliative care. Lancet, 2020

Fairness

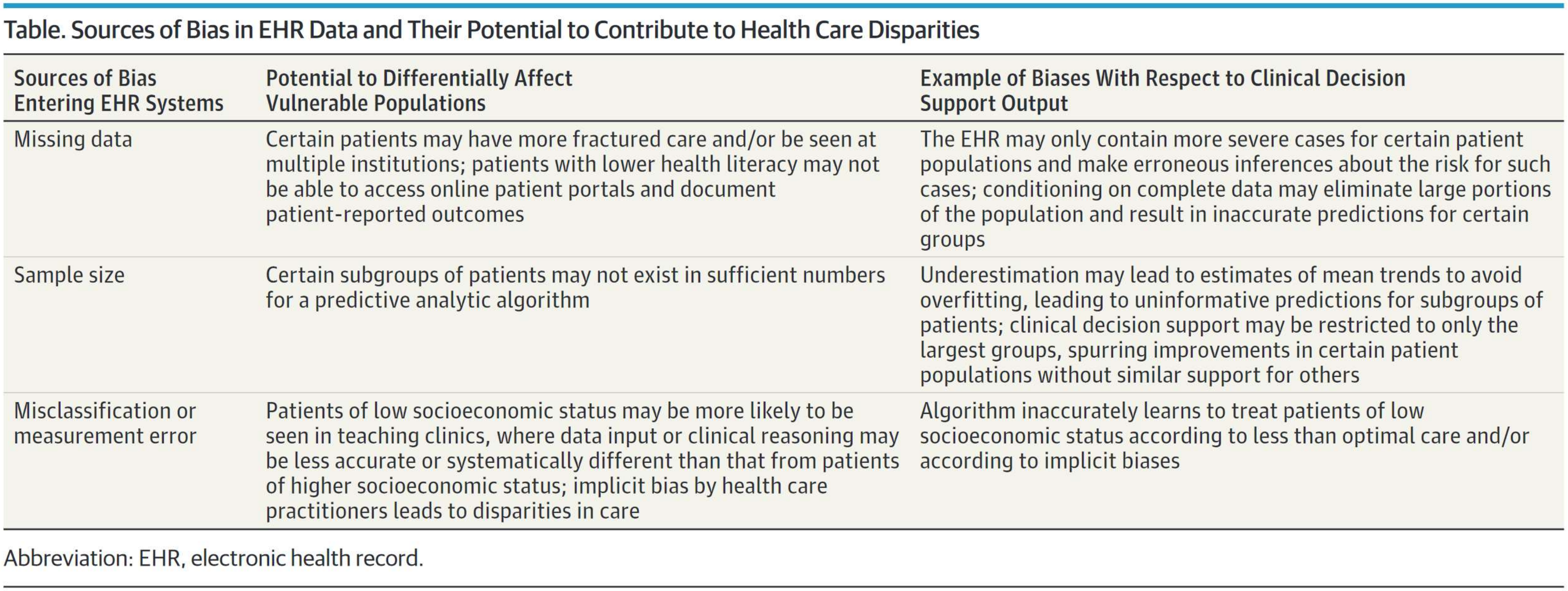


Table. Sources of Bias in EHR Data and Their Potential to Contribute to Health Care Disparities

Sources of Bias Entering EHR Systems	Potential to Differentially Affect Vulnerable Populations	Example of Biases With Respect to Clinical Decision Support Output
Missing data	Certain patients may have more fractured care and/or be seen at multiple institutions; patients with lower health literacy may not be able to access online patient portals and document patient-reported outcomes	The EHR may only contain more severe cases for certain patient populations and make erroneous inferences about the risk for such cases; conditioning on complete data may eliminate large portions of the population and result in inaccurate predictions for certain groups
Sample size	Certain subgroups of patients may not exist in sufficient numbers for a predictive analytic algorithm	Underestimation may lead to estimates of mean trends to avoid overfitting, leading to uninformative predictions for subgroups of patients; clinical decision support may be restricted to only the largest groups, spurring improvements in certain patient populations without similar support for others
Misclassification or measurement error	Patients of low socioeconomic status may be more likely to be seen in teaching clinics, where data input or clinical reasoning may be less accurate or systematically different than that from patients of higher socioeconomic status; implicit bias by health care practitioners leads to disparities in care	Algorithm inaccurately learns to treat patients of low socioeconomic status according to less than optimal care and/or according to implicit biases

Abbreviation: EHR, electronic health record.

Gianfrancesco et al. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med, 2018

How to improve interpretability?

- I.e., how to improve the ability of users to verify requirements?
  - Depends on the requirements!

- AI models are often made more interpretable by using a second model to generate **explanations**
- Another way to increase interpretability is using more **transparent** learning methods
  - E.g., regression, decision trees

Types of explanations

- Local explanations: Why did the model predict Y_i given X_i?
  - E.g. because of these values/pixels
  - E.g. because *i* resembles these cases

- Global explanations: What does the model rely on among X to predict Y?
  - E.g. these features are the most important for accuracy

Explanations need to be interpretable

- Simply using an explanation generating method does not make the model interpretable
- No point if clinicians don’t understand the explanation
- Extensive feature engineering can make this challenging

Avati et al. Improving palliative care with deep learning. BMC Medical Informatics and Decision Making, 2018

Explanations need to be interpretable

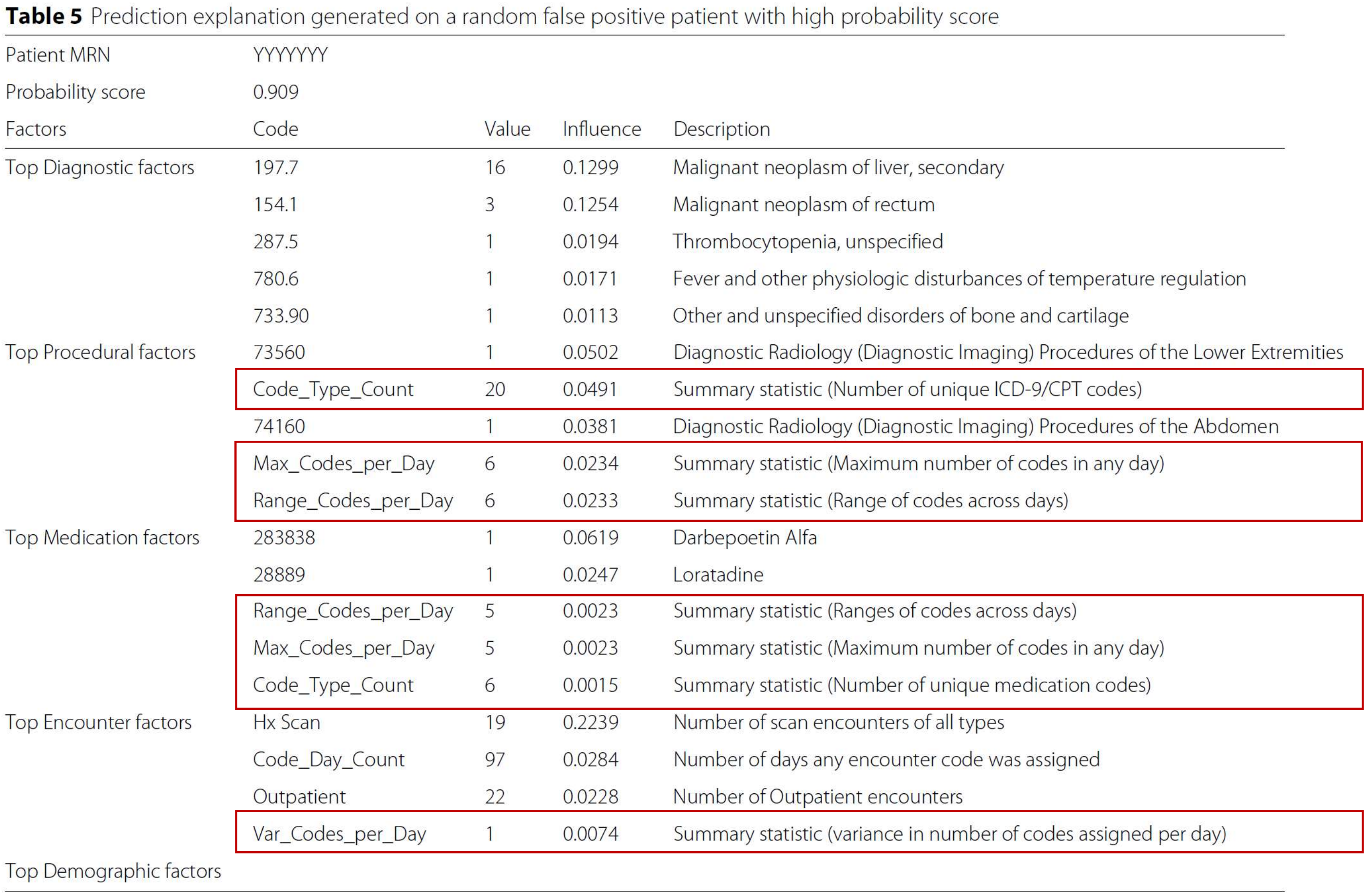
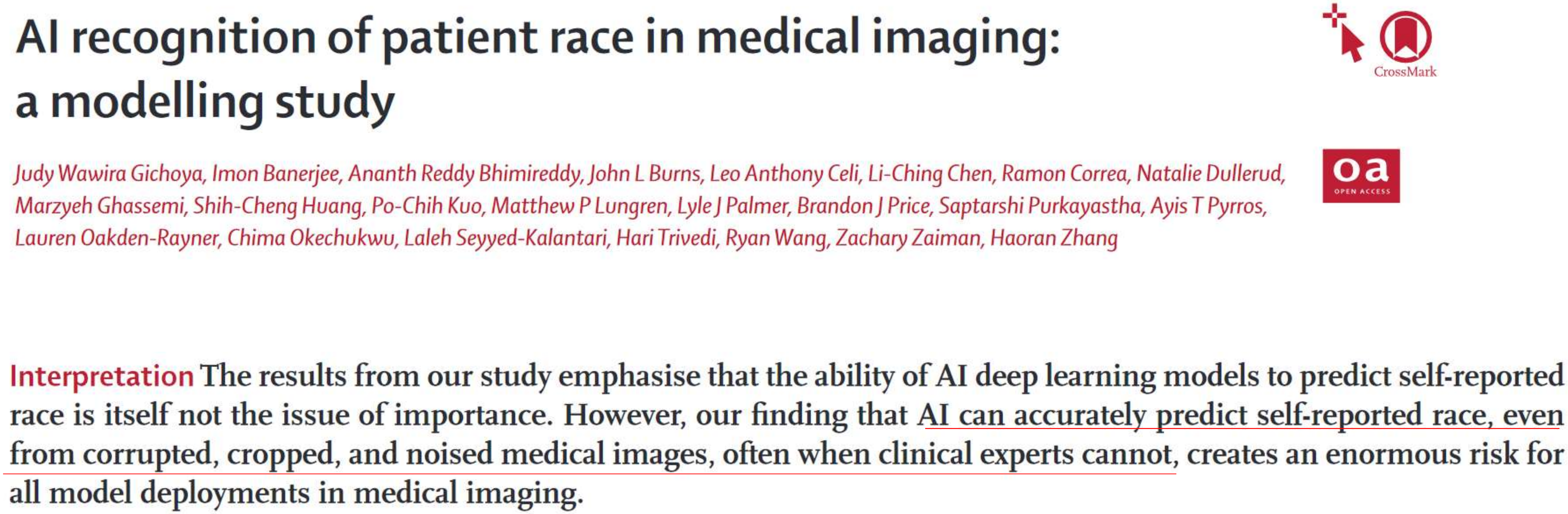


Table 5 Prediction explanation generated on a random false positive patient with high probability score

Patient MRN	YYYYYYY			
Probability score	0.909			
Factors	Code	Value	Influence	Description
Top Diagnostic factors	197.7	16	0.1299	Malignant neoplasm of liver, secondary
	154.1	3	0.1254	Malignant neoplasm of rectum
	287.5	1	0.0194	Thrombocytopenia, unspecified
	780.6	1	0.0171	Fever and other physiologic disturbances of temperature regulation
	733.90	1	0.0113	Other and unspecified disorders of bone and cartilage
Top Procedural factors	73560	1	0.0502	Diagnostic Radiology (Diagnostic Imaging) Procedures of the Lower Extremities
	Code_Type_Count	20	0.0491	Summary statistic (Number of unique ICD-9/CPT codes)
	74160	1	0.0381	Diagnostic Radiology (Diagnostic Imaging) Procedures of the Abdomen
	Max_Codes_per_Day	6	0.0234	Summary statistic (Maximum number of codes in any day)
	Range_Codes_per_Day	6	0.0233	Summary statistic (Range of codes across days)
Top Medication factors	283838	1	0.0619	Darbepoetin Alfa
	28889	1	0.0247	Loratadine
	Range_Codes_per_Day	5	0.0023	Summary statistic (Ranges of codes across days)
	Max_Codes_per_Day	5	0.0023	Summary statistic (Maximum number of codes in any day)
	Code_Type_Count	6	0.0015	Summary statistic (Number of unique medication codes)
Top Encounter factors	Hx Scan	19	0.2239	Number of scan encounters of all types
	Code_Day_Count	97	0.0284	Number of days any encounter code was assigned
	Outpatient	22	0.0228	Number of Outpatient encounters
	Var_Codes_per_Day	1	0.0074	Summary statistic (variance in number of codes assigned per day)
Top Demographic factors				

Limits to explanations



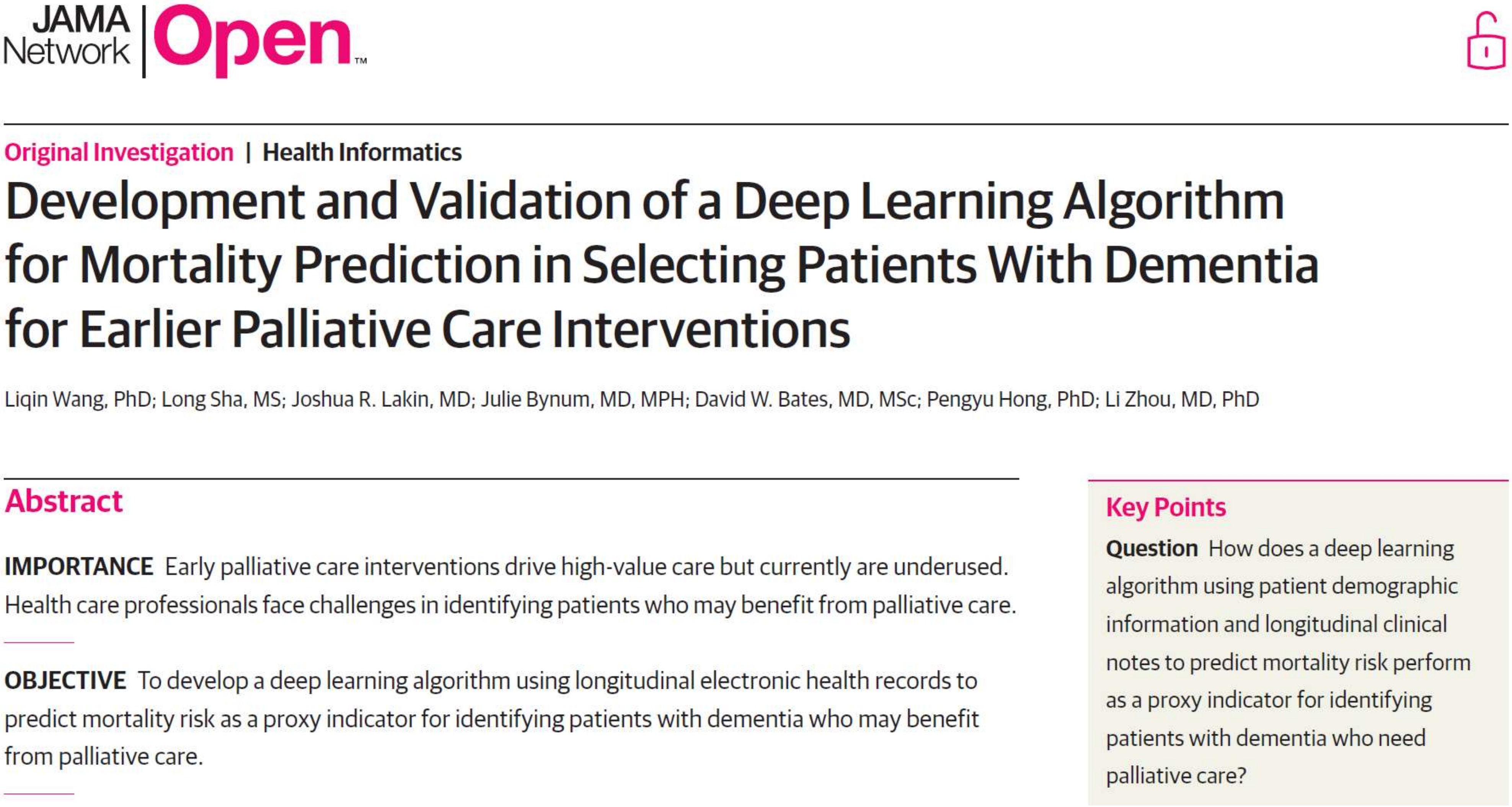
AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

Interpretation The results from our study emphasise that the ability of AI deep learning models to predict self-reported race is itself not the issue of importance. However, our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging.

AI models need to be useful

- An AI model might be applicable, accurate and interpretable, but that doesn’t mean it’ll be useful



JAMA Network Open

Original Investigation | Health Informatics

Development and Validation of a Deep Learning Algorithm for Mortality Prediction in Selecting Patients With Dementia for Earlier Palliative Care Interventions

Liqin Wang, PhD; Long Sha, MS; Joshua R. Lakin, MD; Julie Bynum, MD, MPH; David W. Bates, MD, MSc; Pengyu Hong, PhD; Li Zhou, MD, PhD

Abstract

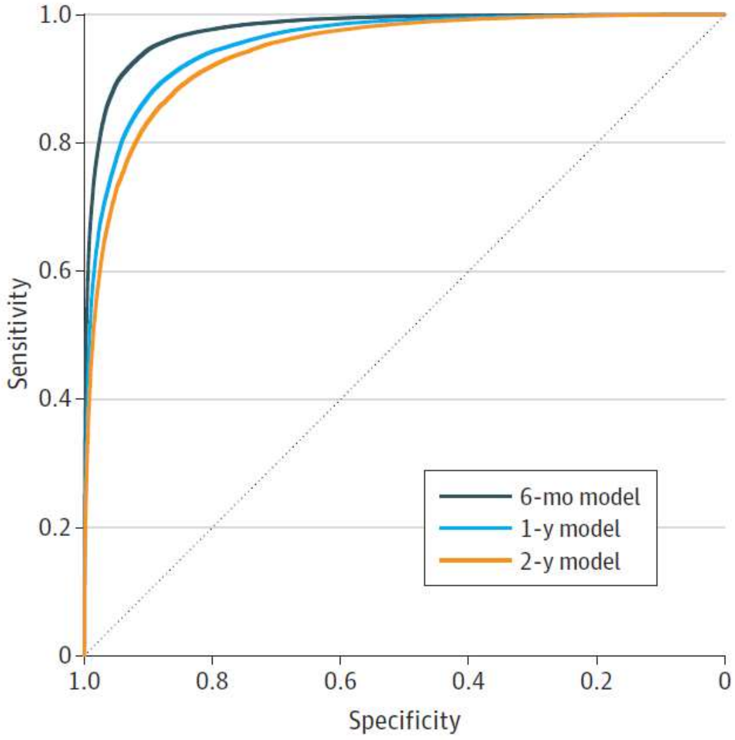
IMPORTANCE Early palliative care interventions drive high-value care but currently are underused. Health care professionals face challenges in identifying patients who may benefit from palliative care.

OBJECTIVE To develop a deep learning algorithm using longitudinal electronic health records to predict mortality risk as a proxy indicator for identifying patients with dementia who may benefit from palliative care.

Key Points

Question How does a deep learning algorithm using patient demographic information and longitudinal clinical notes to predict mortality risk perform as a proxy indicator for identifying patients with dementia who need palliative care?

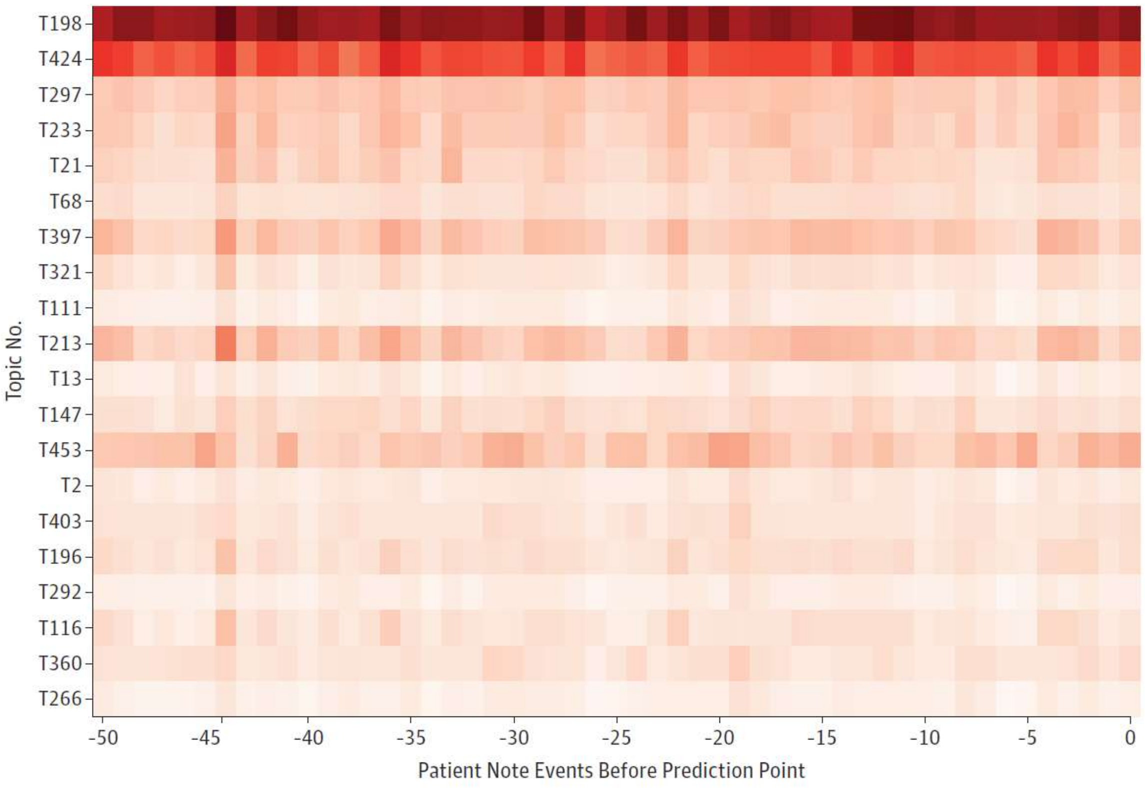
Figure 2. Receiver Operating Characteristic Curves of the Deep Learning Models in Predicting Patient Mortality



In a validation data set of 2692 patients with Alzheimer disease and related dementia, the deep learning–based models showed high note events–level classification of 6-month, 1-year, and 2-year mortality, achieving areas under the receiver operating characteristic curve of 0.978 (95% CI, 0.977-0.978) for the 6-month model, 0.956 (95% CI, 0.955-0.956) for the 1-year model, and 0.943 (95% CI, 0.942-0.944) for the 2-year model.

Figure 3. Topic Attention Heatmap and Corresponding Note Events Predicting 2-Year Mortality

A Topic attention heat map



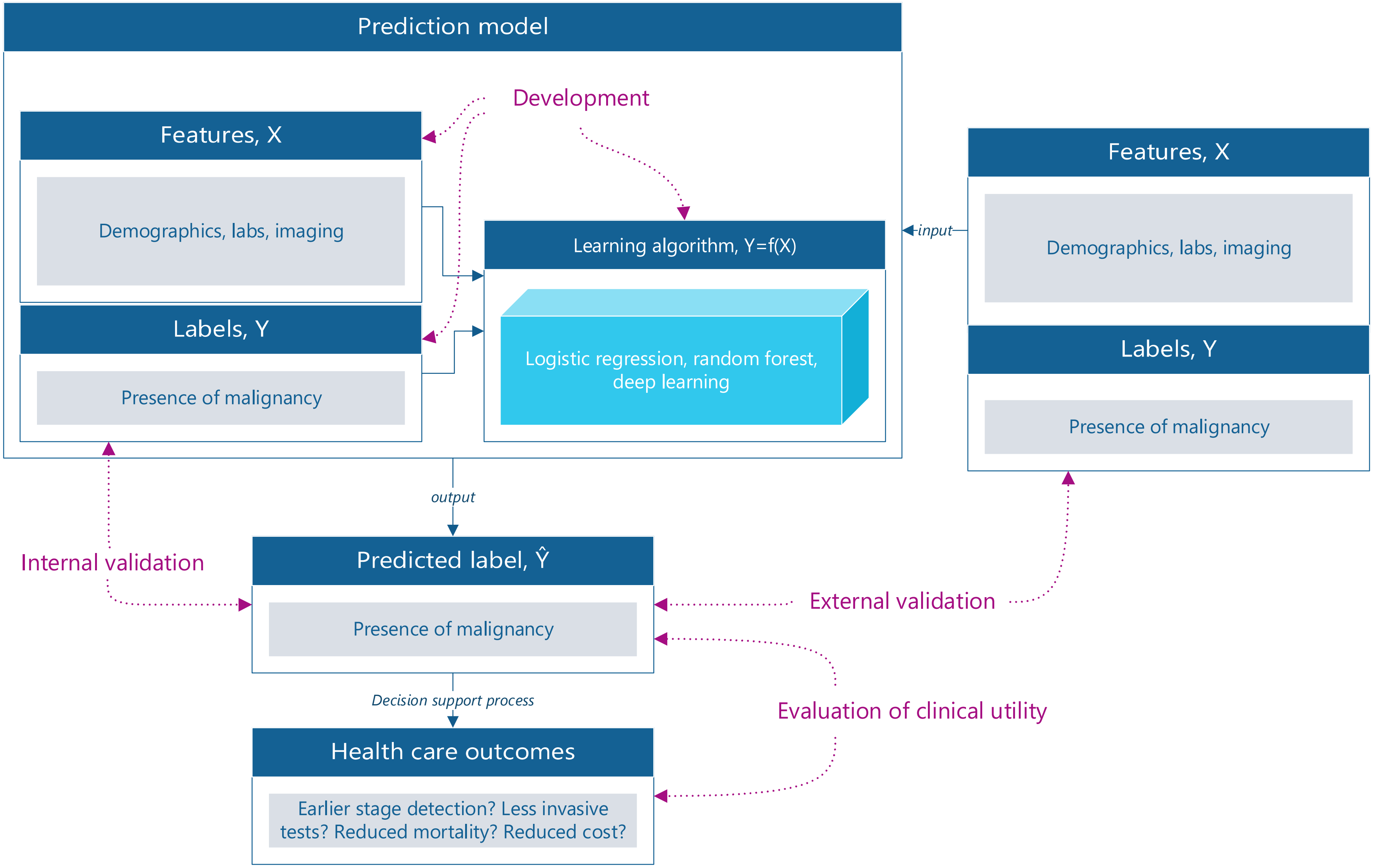
B Top 10 probable words for each topic

Topic No.	Top 10 Words of the Topic
T198	Care hospice family comfort palliative DNI DNR PRN goal morphine
T424	Status dementia unable mental baseline eye command nurse verbal alter
T297	Agitation agitate dementia Seroquel delirium Haldol continue Zyprexa PRN sitter
T233	Cancer lung metastatic disease chemotherapy oncology cycle radiation cell show

The most **predictive** terms correspond to the most **useless** scenarios for supporting decisions

AI models need to be useful

- What is the **impact** of the model on **clinical outcomes**?
- Outcomes are domain-specific
  - Reducing mortality
  - Reducing morbidity
  - Reducing costs
  - Increasing quality of life
  - Increasing efficiency
  - …
- Gold standard: clinical trial ($$$) after deployment ($$$)
- Clinical utility can (and should) be evaluated before implementation


Prediction model
Development
Features, X
Demographics, labs, imaging
Labels, Y
Presence of malignancy
Learning algorithm, Y=f(X)
Logistic regression, random forest, deep learning
input
Features, X
Demographics, labs, imaging
Labels, Y
Presence of malignancy
output
Internal validation
Predicted label, Ŷ
Presence of malignancy
External validation
Decision support process
Evaluation of clinical utility
Health care outcomes
Earlier stage detection? Less invasive tests? Reduced mortality? Reduced cost?

Methods for evaluating clinical utility *in-silico*

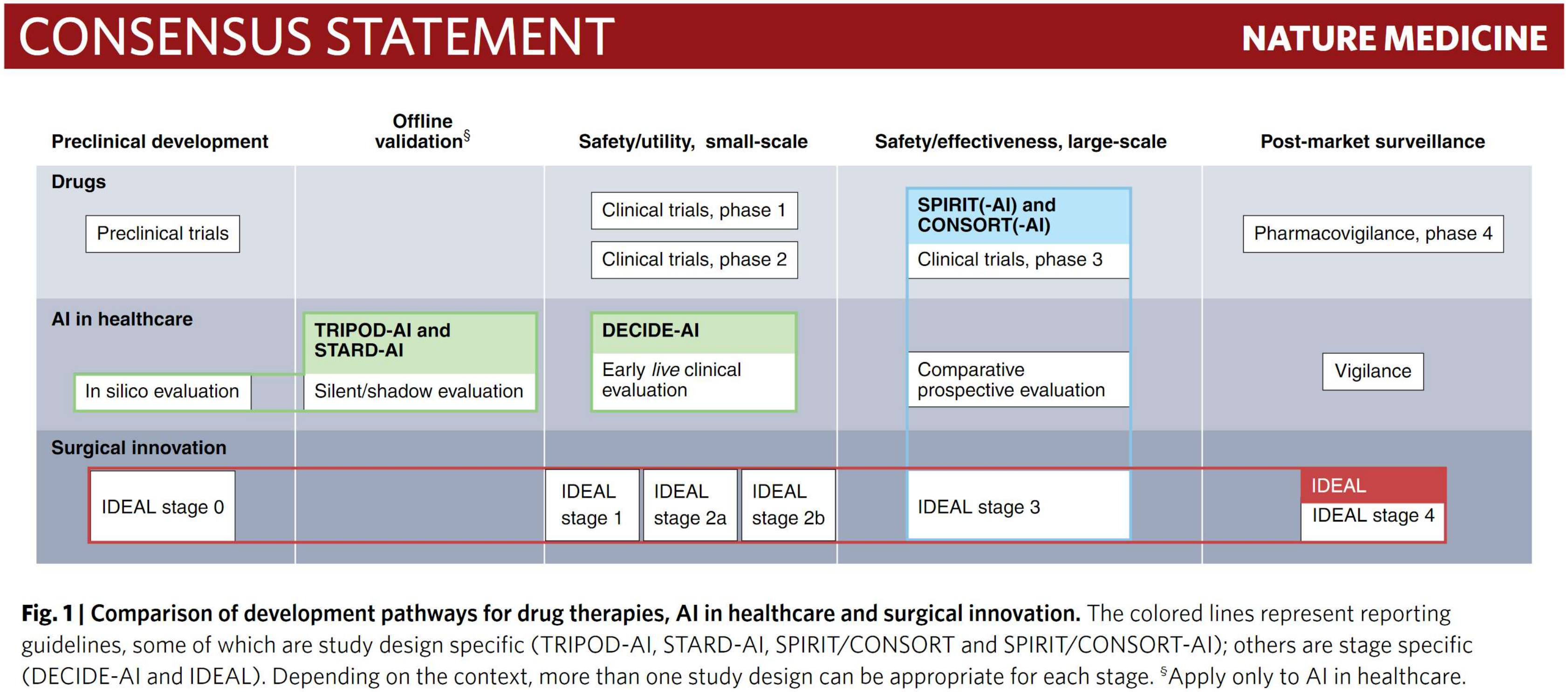
- Decision curve analysis
- Number needed to benefit
- Expected risk difference
- Goal is to evaluate the *expected* value (benefit – cost) of the model relative to alternatives
- Need to be tailored to the use case (need clinician input)
- Better metrics for communicating the importance of a model
  - Model has an AUC of 0.95 vs Model is expected to improve outcome X by %
- Better metrics for justifying deployment costs
- N.B: does not replace evaluation of impact after deployment

Essential elements of evaluating utility

- What is the **problem** the model addresses?
  - E.g. Diagnostic accuracy for cancer on screening test is suboptimal
- What are the **beneficial outcomes** when the model is right?
  - E.g. Detecting a cancer early
- What are the **costly outcomes** when the model is wrong?
  - E.g. Biopsy of a benign lesion
- What is (are) the **model threshold**(s) for prompting action?
  - E.g. How high should the probability of cancer be to proceed with biopsy?
- What are the **alternative** strategies?

Hunink et al. Decision Making in Health and Medicine, Cambridge University Press, 2014

Recent reporting guidelines for evaluating utility/safety in small scale deployment



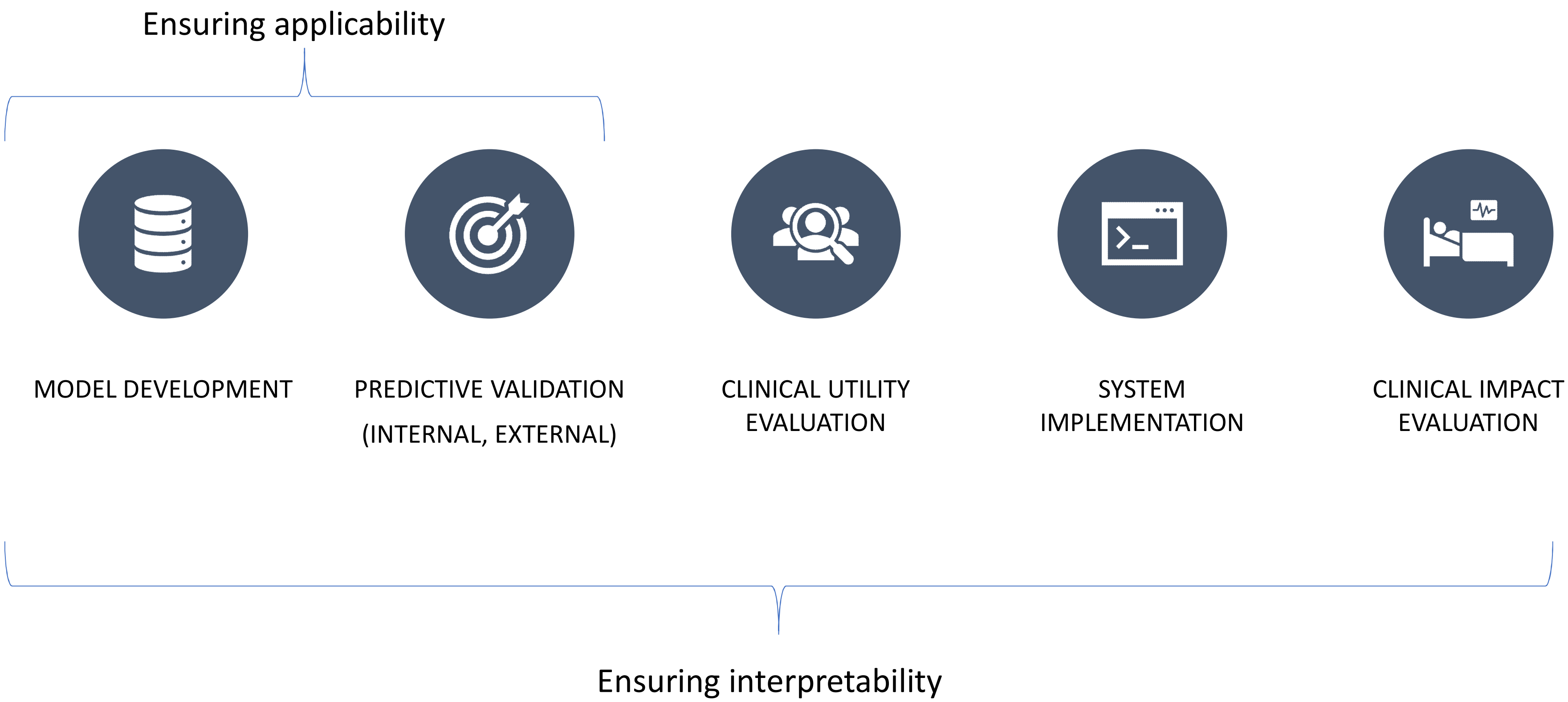
CONSENSUS STATEMENT NATURE MEDICINE

	Preclinical development	Offline validation§	Safety/utility, small-scale	Safety/effectiveness, large-scale	Post-market surveillance
Drugs	Preclinical trials		Clinical trials, phase 1 Clinical trials, phase 2	SPIRIT(-AI) and CONSORT(-AI) Clinical trials, phase 3	Pharmacovigilance, phase 4
AI in healthcare	In silico evaluation	TRIPOD-AI and STARD-AI Silent/shadow evaluation	DECIDE-AI Early *live* clinical evaluation	Comparative prospective evaluation	Vigilance
Surgical innovation	IDEAL stage 0		IDEAL stage 1 IDEAL stage 2a IDEAL stage 2b	IDEAL stage 3	IDEAL IDEAL stage 4

Fig. 1 | Comparison of development pathways for drug therapies, AI in healthcare and surgical innovation. The colored lines represent reporting guidelines, some of which are study design specific (TRIPOD-AI, STARD-AI, SPIRIT/CONSORT and SPIRIT/CONSORT-AI); others are stage specific (DECIDE-AI and IDEAL). Depending on the context, more than one study design can be appropriate for each stage. §Apply only to AI in healthcare.

Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, Nature Medicine, 2022

Byte-to-bedside process



Ensuring applicability

MODEL DEVELOPMENT

PREDICTIVE VALIDATION

(INTERNAL, EXTERNAL)

CLINICAL UTILITY
EVALUATION

SYSTEM
IMPLEMENTATION

CLINICAL IMPACT
EVALUATION

Ensuring interpretability

Conclusion

- Is my model applicable, valid, interpretable and useful?
  - Depends on the use case!
- Clinical stakeholder perspective:
  - Can it be applied locally in real-time?
  - When it is applied, will it be good at what it is trained to do?
  - If it makes good predictions, can I use/trust it to make decisions?
  - If I use/trust it, how will it impact clinical outcomes?