Stanford
MEDICINE

Frontiers of AI in Medical Imaging:

Overcoming Current Challenges and Moving Beyond Classification

Daniel L. Rubin (MD) and Imon Banerjee (PhD)
Department of Biomedical Data Science and Radiology
Stanford School of Medicine

Acknowledgements

Funding Support

NCI QIN grants U01CA142555,1U01CA190214, 1U01CA187947
Stanford-AstraZeneca Collaboration Grant
NVIDIA Academic Hardware Grant Program
Stanford Philips and GE BlueSky

Outline

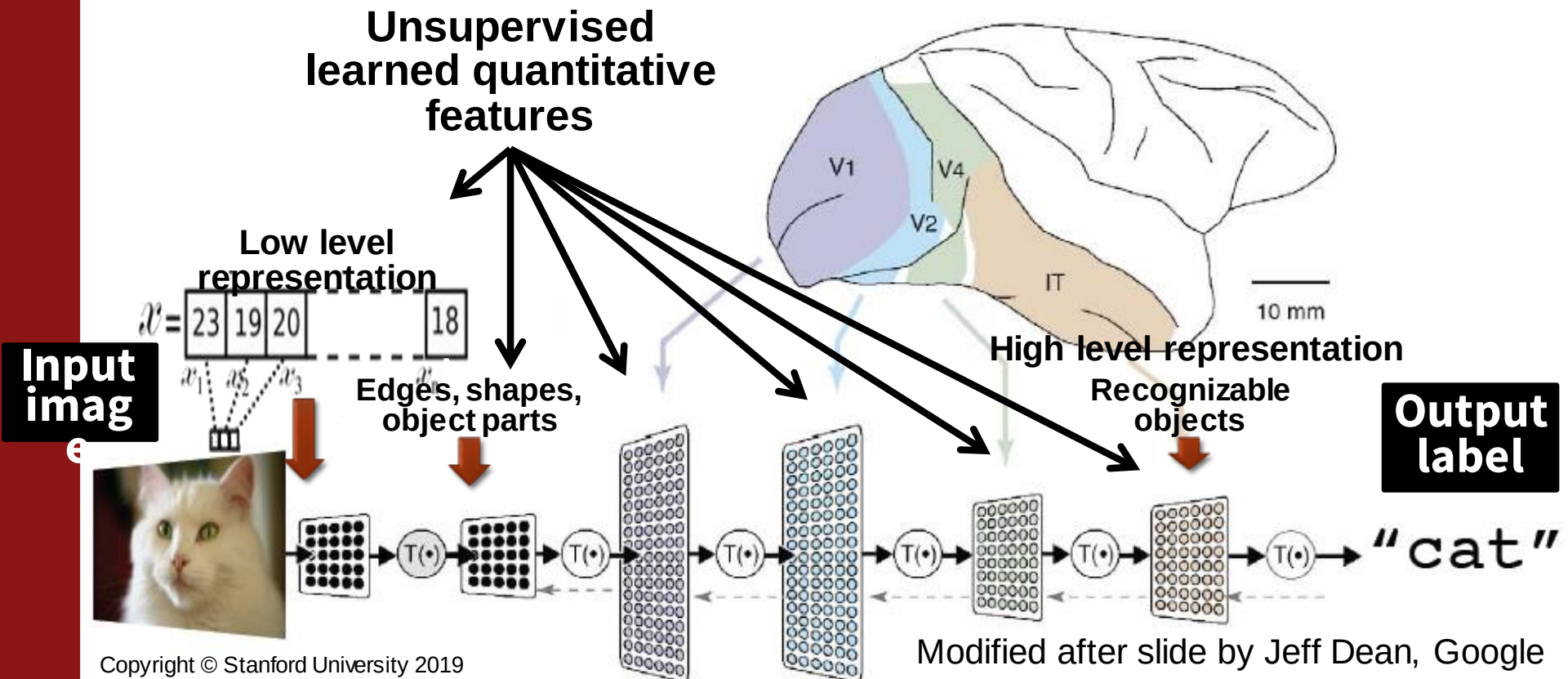
1. Need for image interpretation beyond image classification
2. Integrating multiple data types with images
3. Making AI clinical predictions and providing explanation
4. Evaluation of AI algorithms

Outline

1. Need for image interpretation beyond image classification
2. Integrating multiple data types with images
3. Making AI clinical predictions and providing explanation
4. Evaluation of AI algorithms

Deep learning: Image classification

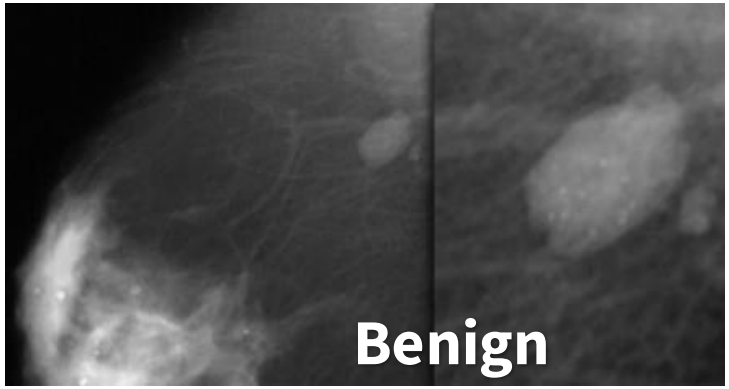
- High-level abstractions of image features hierarchical, non-linear transformations
- Higher-level features (layers) are defined from lower-level ones, and represent higher levels of abstraction
- Most suitable for **classification problems**

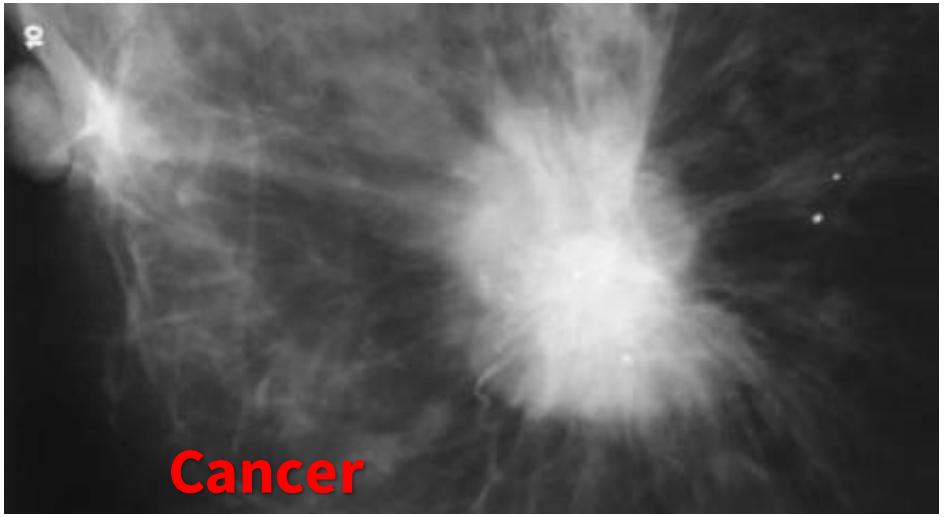


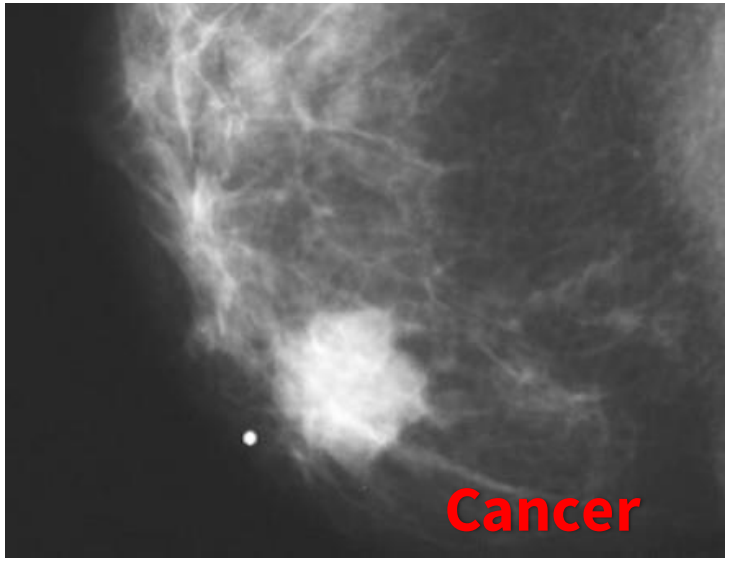
Modified after slide by Jeff Dean, Google

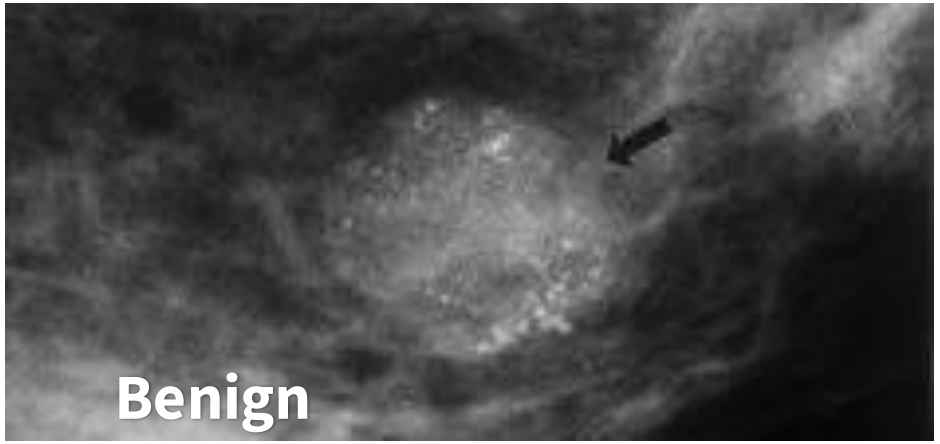
Image classification in medical imaging

"Benign or cancer lesion?"









Stanford University

There are other important medical needs beyond image classification…

Key medical applications ***beyond classification***

1. Disease detection
2. Lesion segmentation
3. Treatment selection
4. Response assessment
5. **Clinical prediction** (of response or future disease)

People (and their diseases) differ…

Stanford University

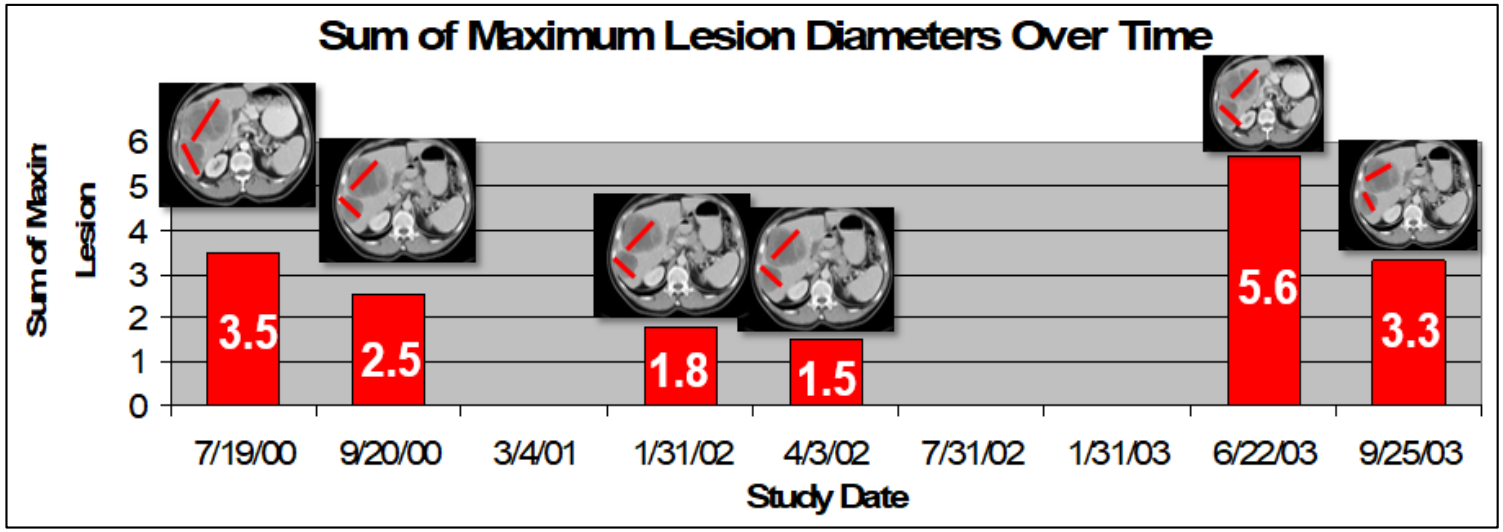
"Precision Medicine"

- Patient care often lacks specificity ("One size fits all" does not usually apply in medicine)
- There are "subtypes" of disease (e.g., many types of "breast cancer" needing specific therapy for each type)
- Precise diagnoses based on "electronic phenotyping" and molecular profiling enables treatments that are tailored to unique characteristics of each patient
- Requires accurate methods of prediction based on disease phenotypes
  - › Key opportunity for Big Data and AI methods

Stanford University

Prediction

- Disease in patients **evolves over time** (longitudinally)



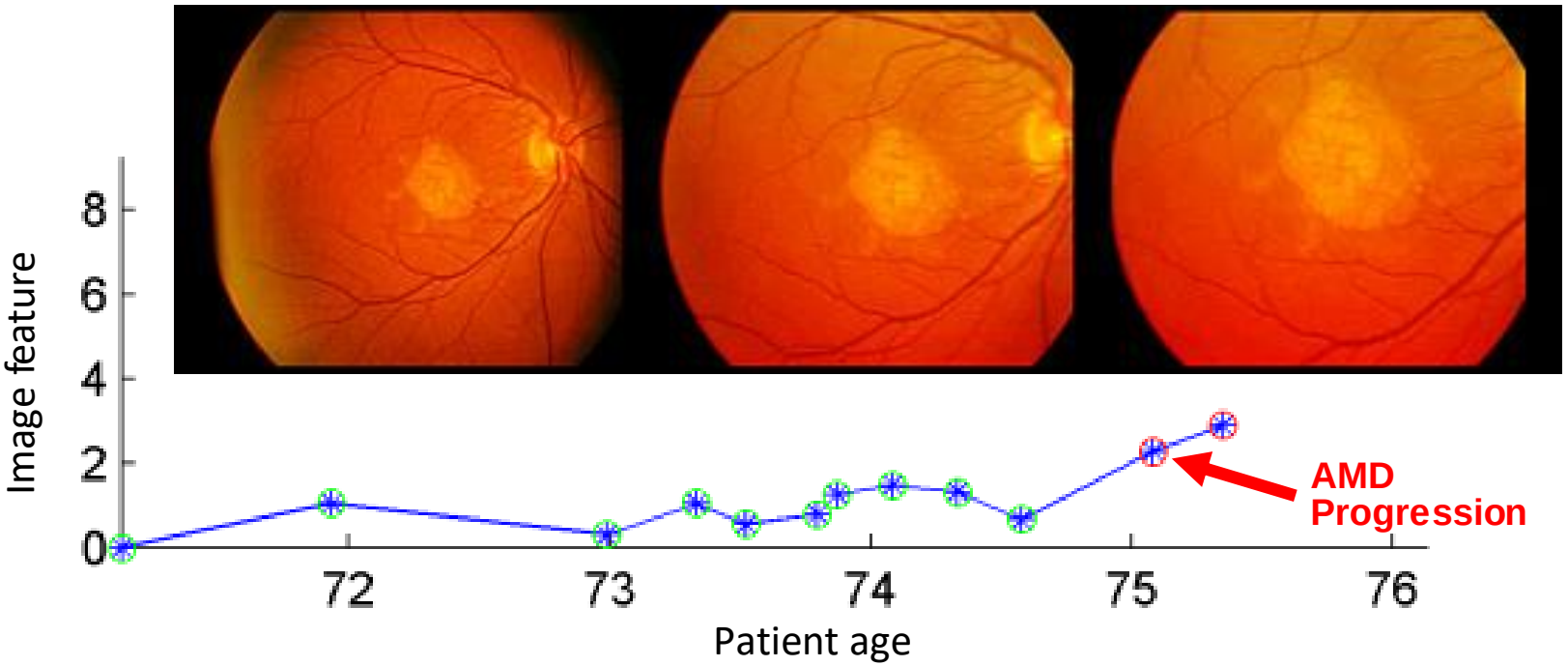
- Patient data (images and text reports/notes) are acquired longitudinally

- **We need prediction models need to account for longitudinal data inputs**

Stanford University

Progression of age related macular degeneration eye disease

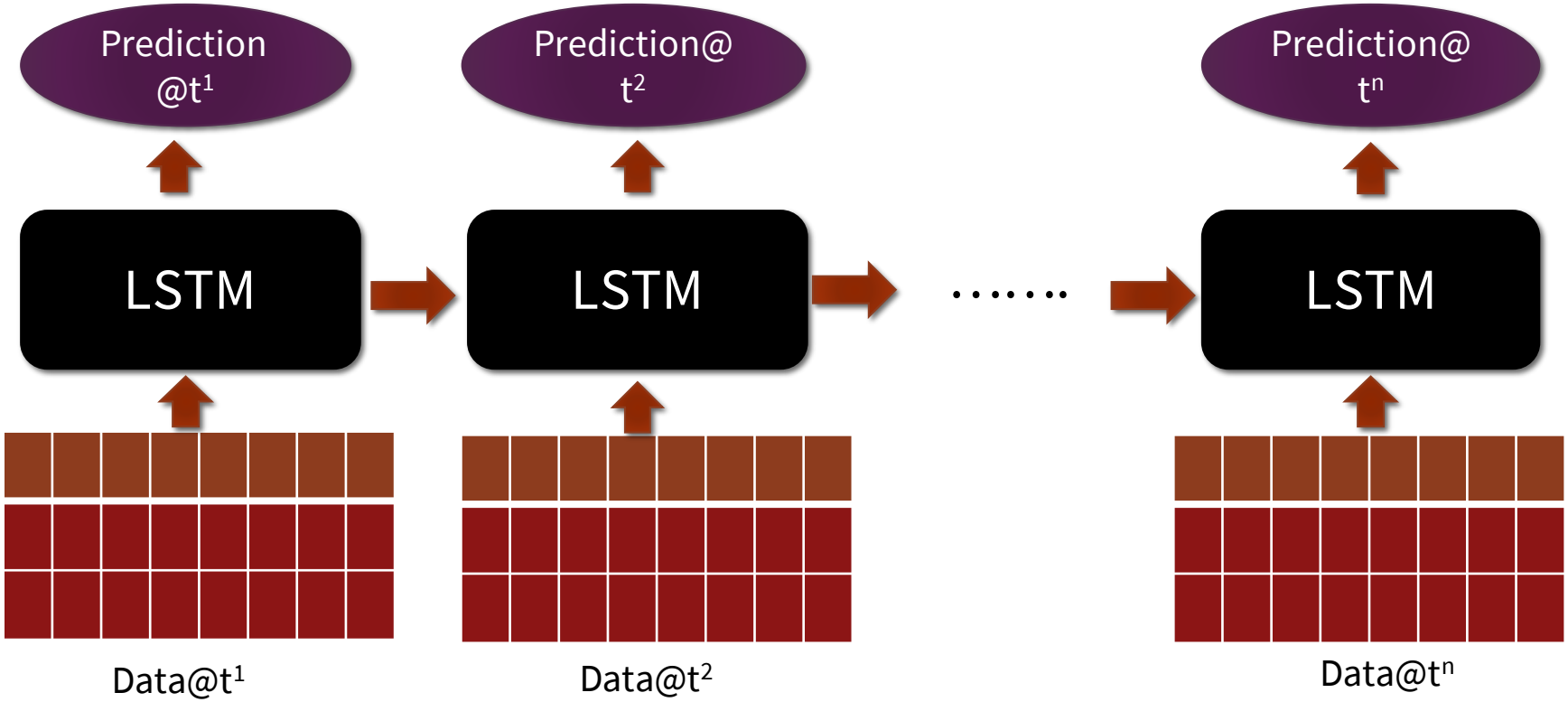
- AMD changes over time
- Some patients progress to wet AMD
- The time to AMD progression is unpredictable



Stanford University

Prediction model (RNN)

- Many-to-many RNN using two-layer one-directional stacked stateful Long short-term memory (LSTM)
- Long-term memory during training encodes information about entire temporal visit sequence
- Short-term memory passes immediate state between successive nodes



Stanford University

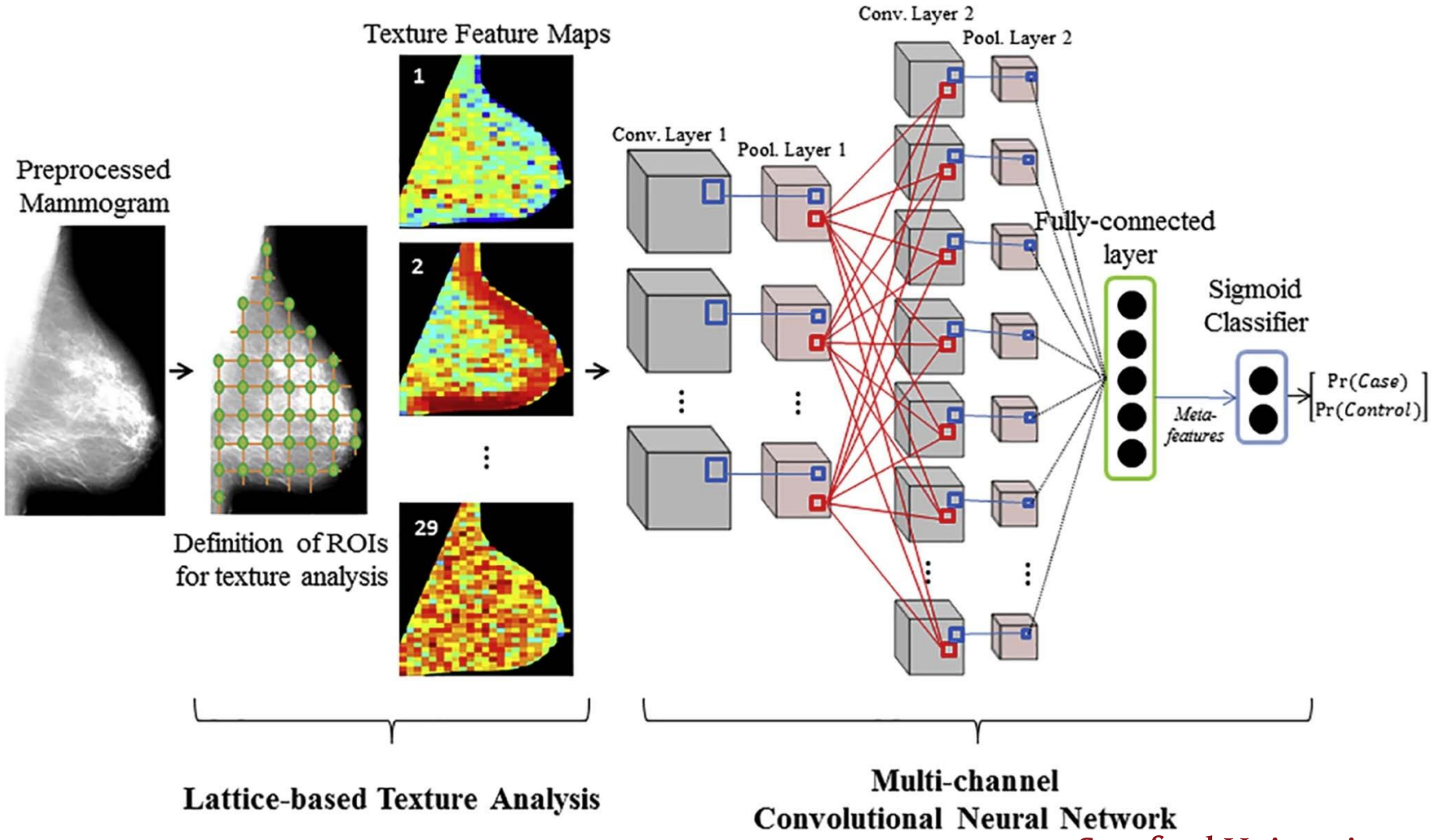
“Precision Health”

- A paradigm shift, focusing on prediction and prevention, rather than relying exclusively on diagnosis and treatment of existing disease
- Prevents or forestalls the development of disease
- Reduces costs and morbidity and improves patient care
- Requires accurate methods of prediction based on monitoring people’s health status
  - Key opportunity for Big Data and AI methods

Stanford University

Deep learning for predicting future cancer risk

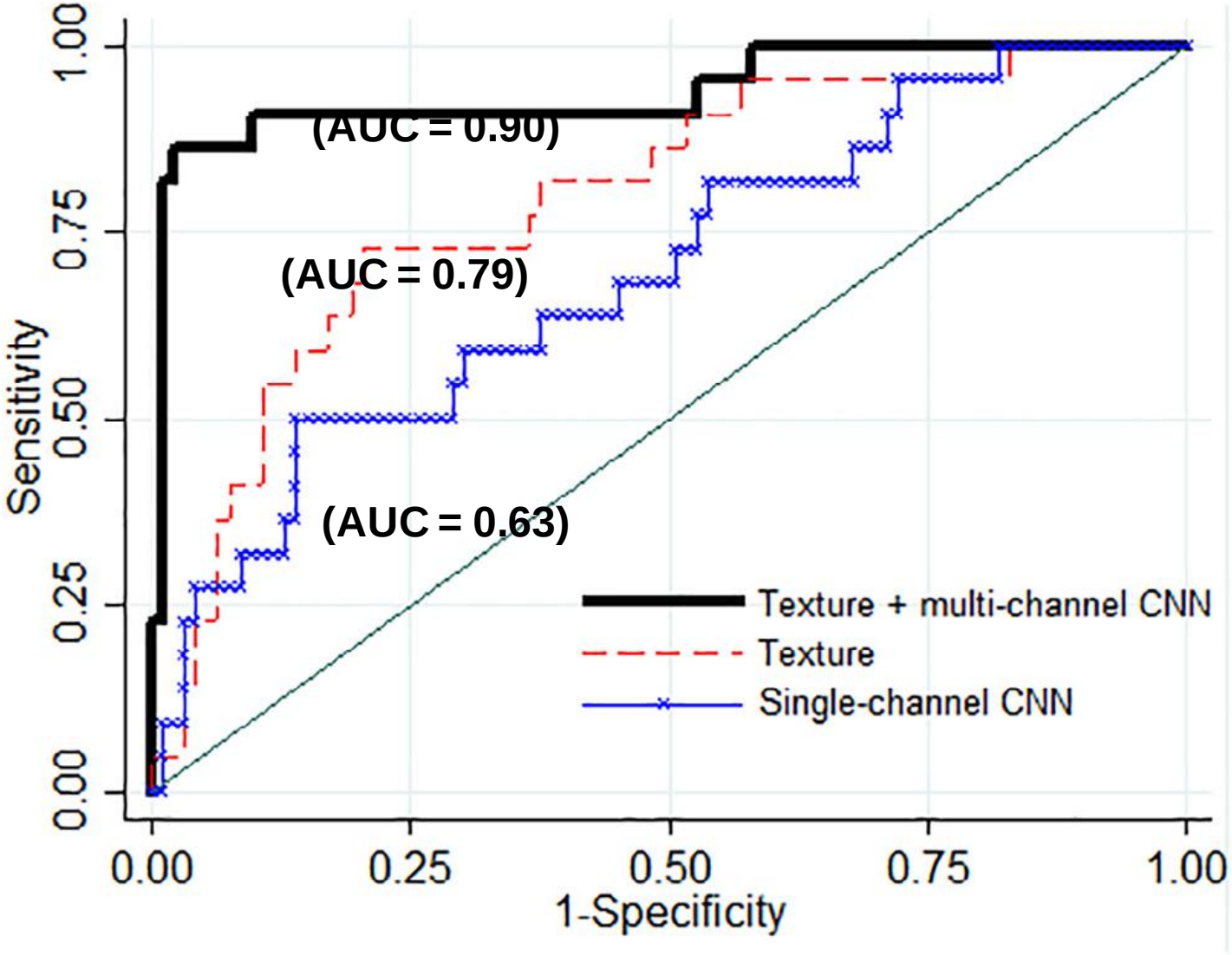
Image texture feature maps preserve discriminative spatially-dependent features and augment data in multi-channel CNN



Acad Radiol 25(8):977-984, 2018

Stanford University

Performance (ROC) of different approaches

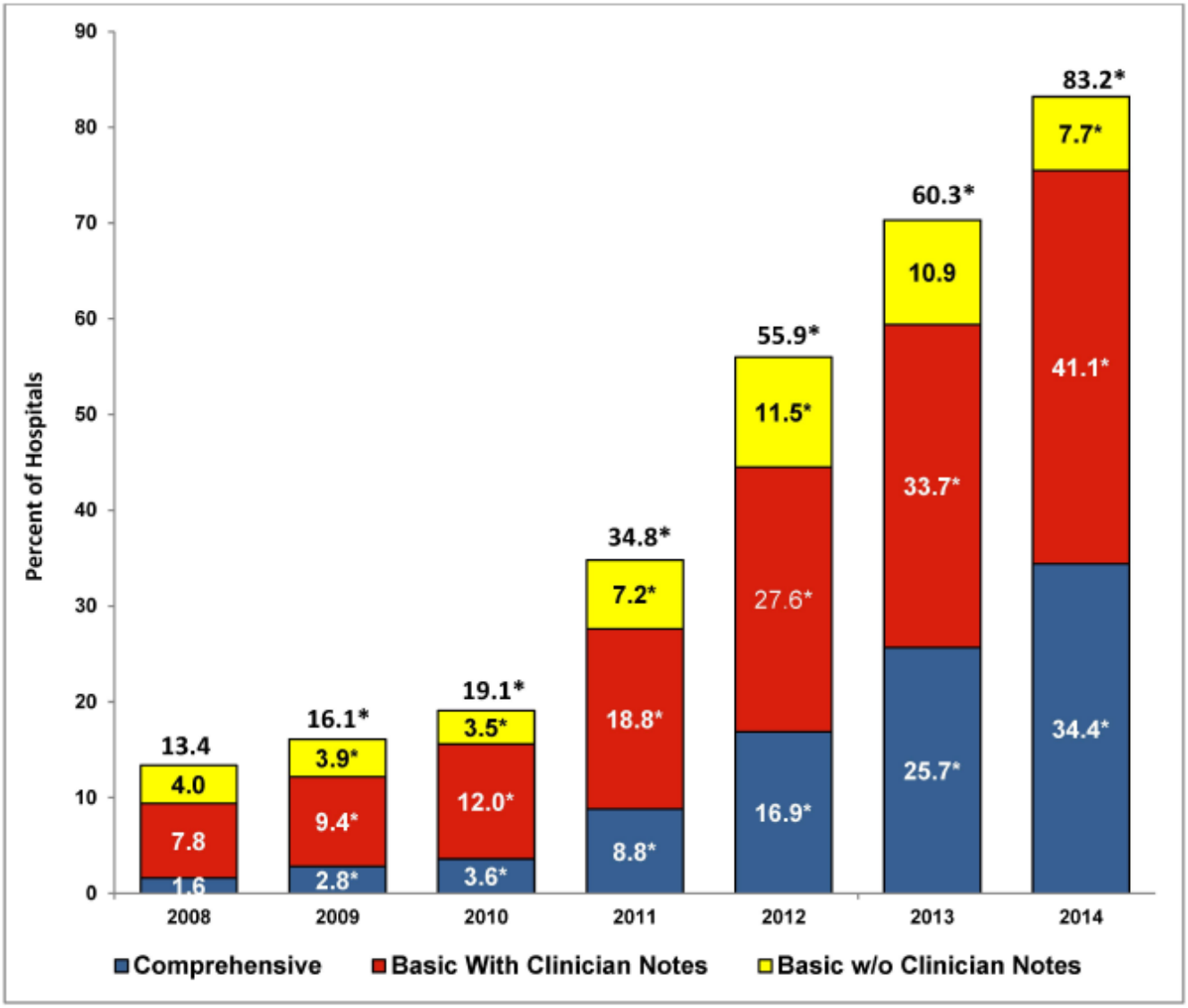


Acad Radiol 25(8):977-984, 2018

Stanford University

Explosion in electronically-accessible medical records data provides opportunity to learn models to help with these prediction problems

Growth in electronic patient data



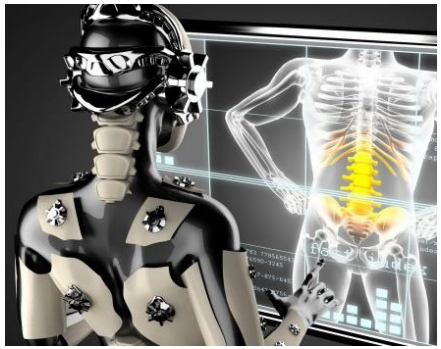
Source: https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf

Stanford University

Outline

1. Need for image interpretation beyond image classification
2. Integrating multiple data types with images
3. Making AI clinical predictions and providing explanation
4. Evaluation of AI algorithms

Integrating various types of data (e.g., images + clinical notes) is needed



Computerized model:

Disease detection

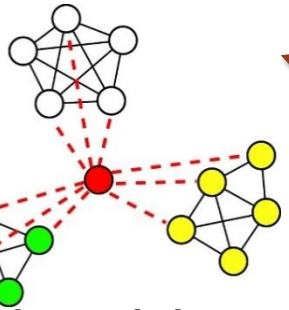
Diagnosis

Treatment response evaluation

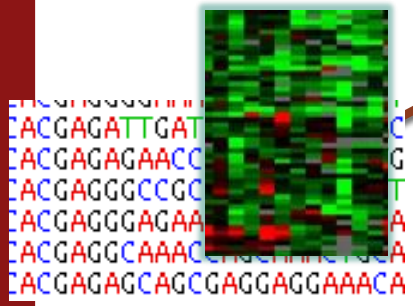
Clinical prediction

Clinical Decision Support

Radiomics and Deep Learning

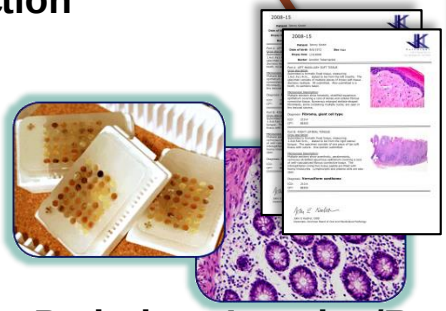


Integration and phenotype extraction

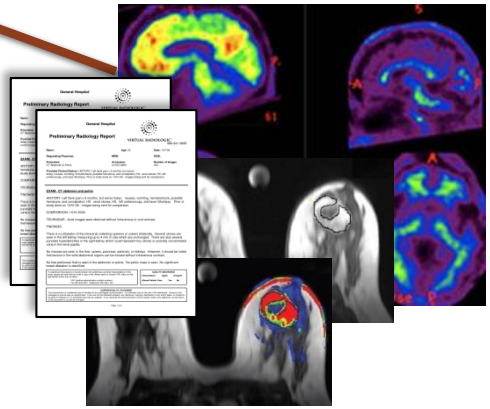


Molecular Characterization

Laboratory and Clinical Testing



Pathology Imaging/Reports



Radiology Imaging and Reports

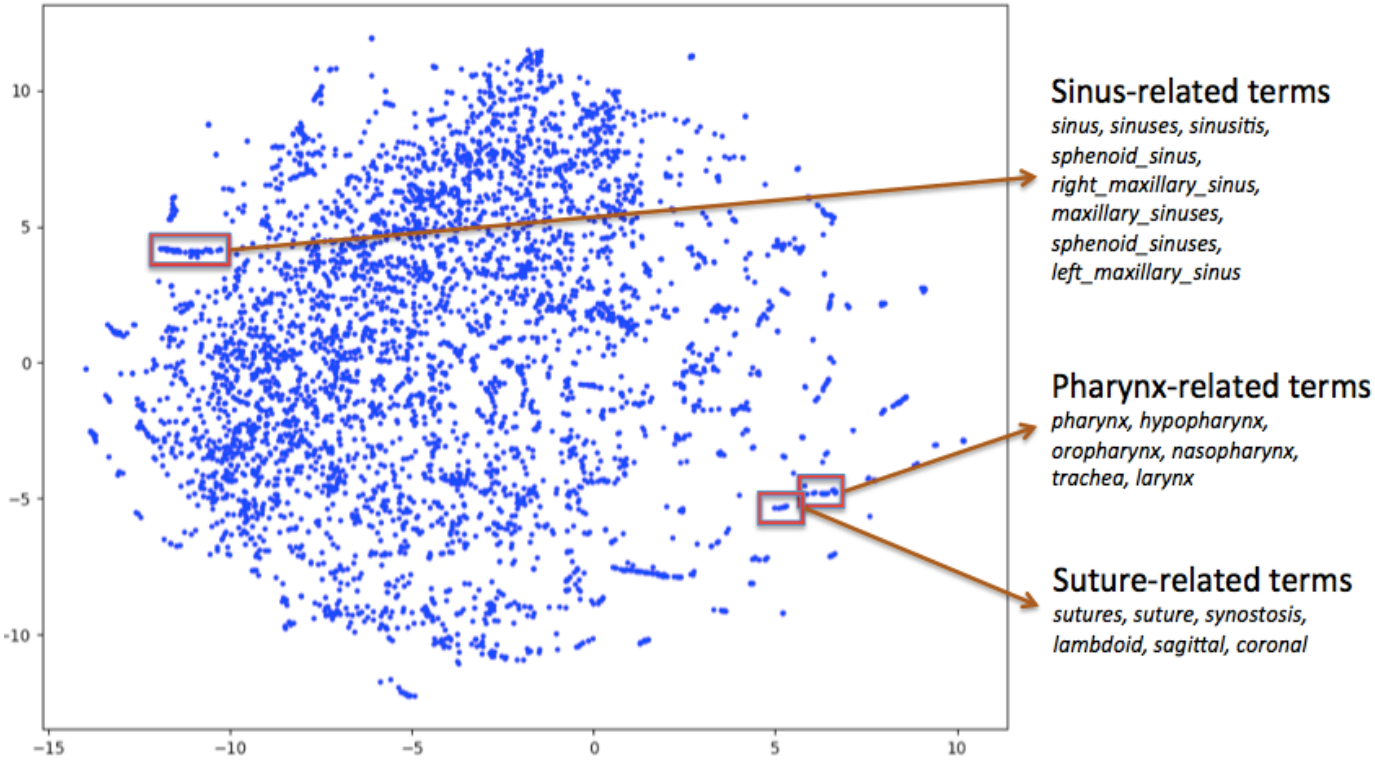
Stanford University

Dealing with narrative text – feature generation

Pathology, radiology report and clinical notes

1. Rule-based and dictionary-based information extraction
2. Bag of word based methods
3. Statistical methods
4. Word embeddings - Word2Vec, GolVe

Identifying core terms from unstructured narrative text

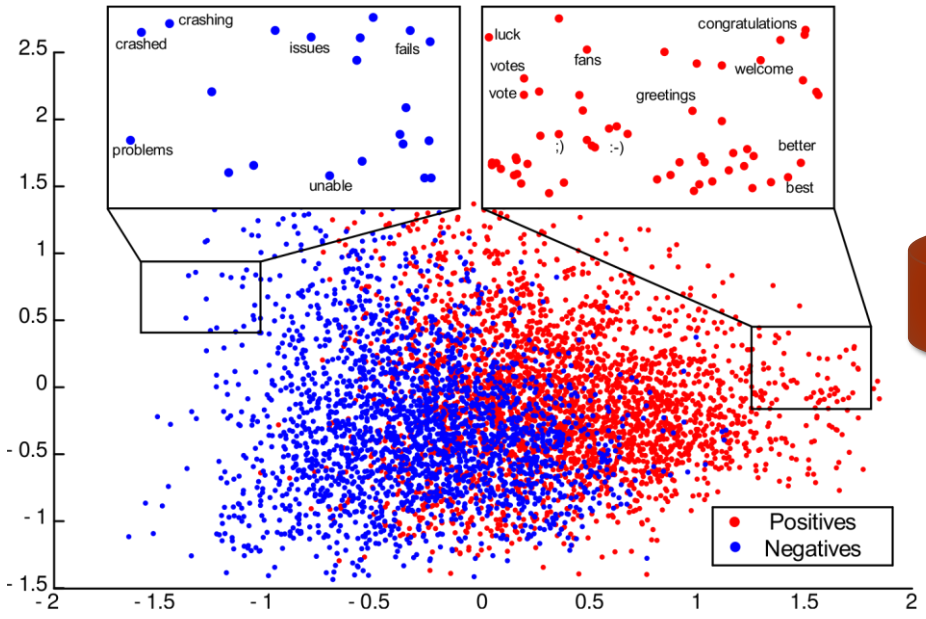


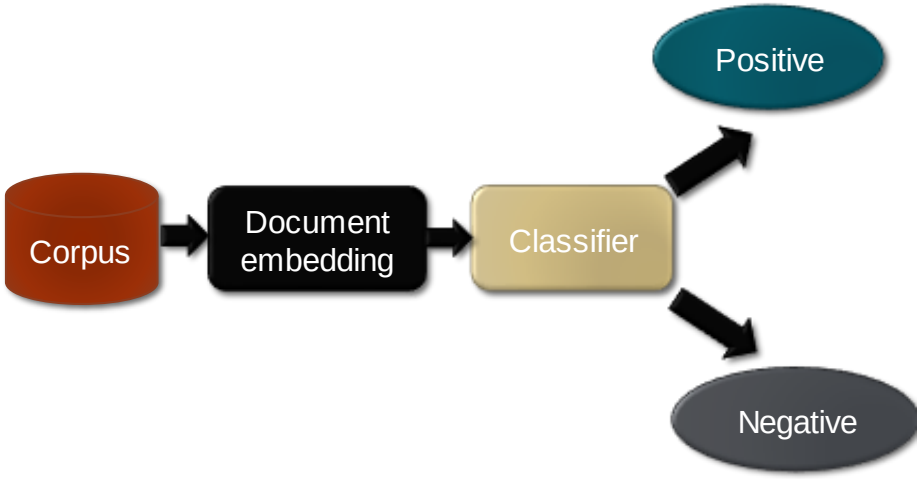
Word embedding using deep learning (4,442 words) projected in two dimensions

Unsupervised deep learning algorithms can discover annotation from texts without the need of supplying specific domain knowledge

Word embedding + classification model

- Stores each word in as a point in space, where it is represented by a vector of fixed number of dimensions.
- Unsupervised, built just by reading huge corpus
- Can be used as features to train a supervised model with a small subset of annotation





Word embedding

Document classification

Mikolov, Distributed representations of words and phrases and their compositionality

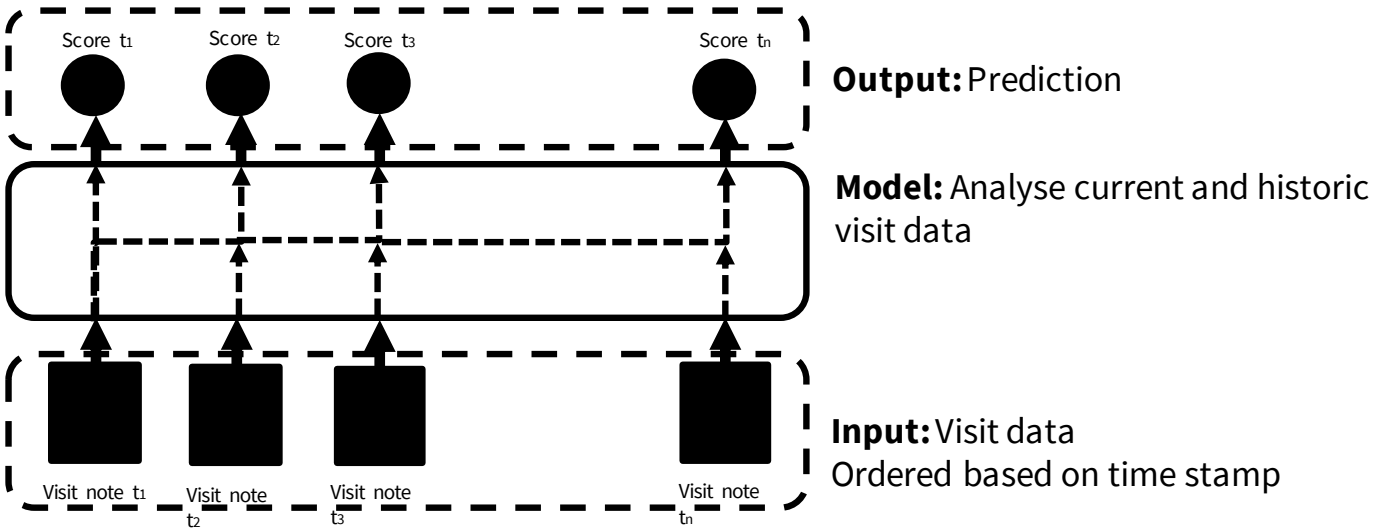
Stanford University

Outline

1. Need for image interpretation beyond image classification
2. Integrating multiple data types with images
3. Making AI clinical predictions and providing explanation
4. Evaluation of AI algorithms

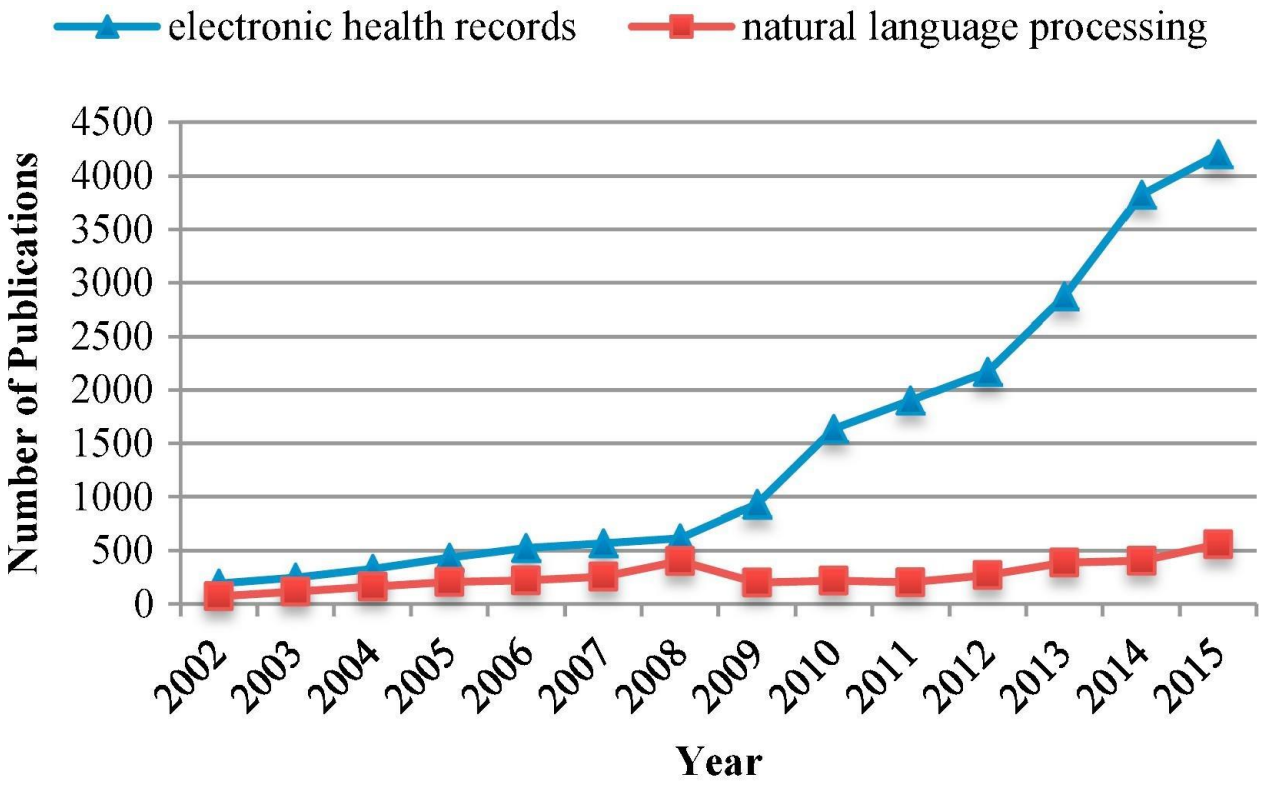
Objective

- Create a dynamic model that takes as input longitudinal visit data ordered according to the date of visits.
- Computes as output a probability of future clinical events for each visit considering the current and all the historic time points.



Stanford University

Under-utilization of NLP in EHR-based research



The number of natural language processing (NLP)-related articles compared to the number of electronic health record (EHR) articles from 2002 through 2015

Yanshan Wang et. al., Clinical information extraction applications: A literature review, JBI 2018

Stanford University

Challenges

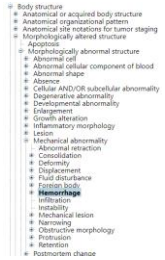
HISTORY: This 69-year-old male returns today immediately upon completion of his renal/bladder ultrasound scan in MMC X-Ray Department. The patient had presented to this office one week ago (XX) with acute onset of lower urinary tract symptoms including nocturia x 5, weakness of his urinary flow and a sensation of incomplete bladder emptying. However, during the course of the next few days, his symptoms gradually resolved. The patient is now relatively asymptomatic from the urologic standpoint having returned to his baseline. Preliminary report concerning his renal/bladder ultrasound scan indicates continued presence of a hypoechogenic focus within the upper pole of the right kidney unchanged from his previous exam in February of this year. Initial bladder volume then was 626 cc with postvoid residual of 104 cc. Today initial bladder volume is 572 cc with postvoid residual of 197 cc. Prostate volume was estimated at 24 cc (February 20XX), increased to 33 cc (today). The patient has been taking Proscar 5 mg daily since July XX. Laboratory results include urinalysis with 1-5 RBCs/HPF, 0-rare WBCs/HPF, hemostix "trace" positive, and leukocyte esterase "negative." Urine culture showed "no growth" on that date. GU EXAM: Trim, generally healthy appearing male with normal, circumcised penis, adequate meatus. Testes are somewhat atrophic and descended bilaterally. Digital rectal exam reveals a prostate gland which is not particularly enlarged (1-2+ enlarged at most), rubbery consistency compressible throughout with smooth surface, intact superior and lateral margins and shallow median groove present. There is no gross nodularity or asymmetry present. IMPRESSION: Acute onset of lower urinary tract symptoms one week ago which proved to be transitory and resolved spontaneously. Urinalysis and urine culture failed to indicate any evidence of urinary infection as the underlying cause of this problem. However, the patient is noted to have rather significant postvoid residual urine (104 cc in February this year and 197 cc today). The prostate gland is modestly enlarged (24 cc in February, 33 cc today) despite ongoing Proscar therapy. However, it is likely the prostate gland would be considerably more enlarged and the patient more consistently symptomatic (urinary outflow obstructive symptoms (had he not been on Proscar during the past five years. PLAN: As the patient's lower urinary tract symptoms have resolved for the most part, it was elected to merely follow him along conservatively for the time being. If the patient develops recurrence of lower urinary tract symptoms, particularly urinary outflow obstructive symptoms, then further urologic intervention may be considered including TUR prostate if indicated. The patient will keep us posted concerning his urologic status.

1. How to extract the relevant sentences?
2. How to determine sentiment of the sentence towards a targeted task?
3. How to label the full notes when multiple sentences reflect different sentiments?

Intelligent Word Embedding (IWE)

NEGEX	Risk
adequate to rule against declined declines demonstrate no ……	and possible and probable cannot exclude cannot r/o chance could lead ……

CleVer Dictionary



Domain-specific Dictionary

Ontocrawler

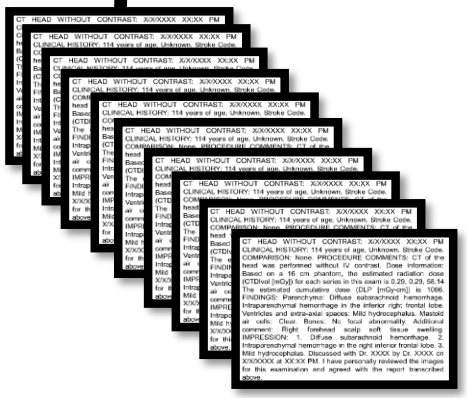
BioPortal

NCBO

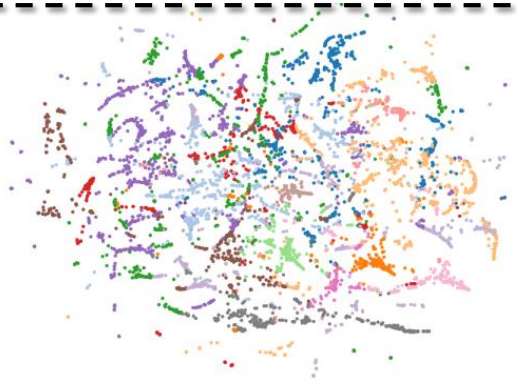
Semantic dictionary generation

Report condenser → Semantic dictionary mapping → Neural word embedding → Context-aware visit vector creation

Unsupervised embedding of reports



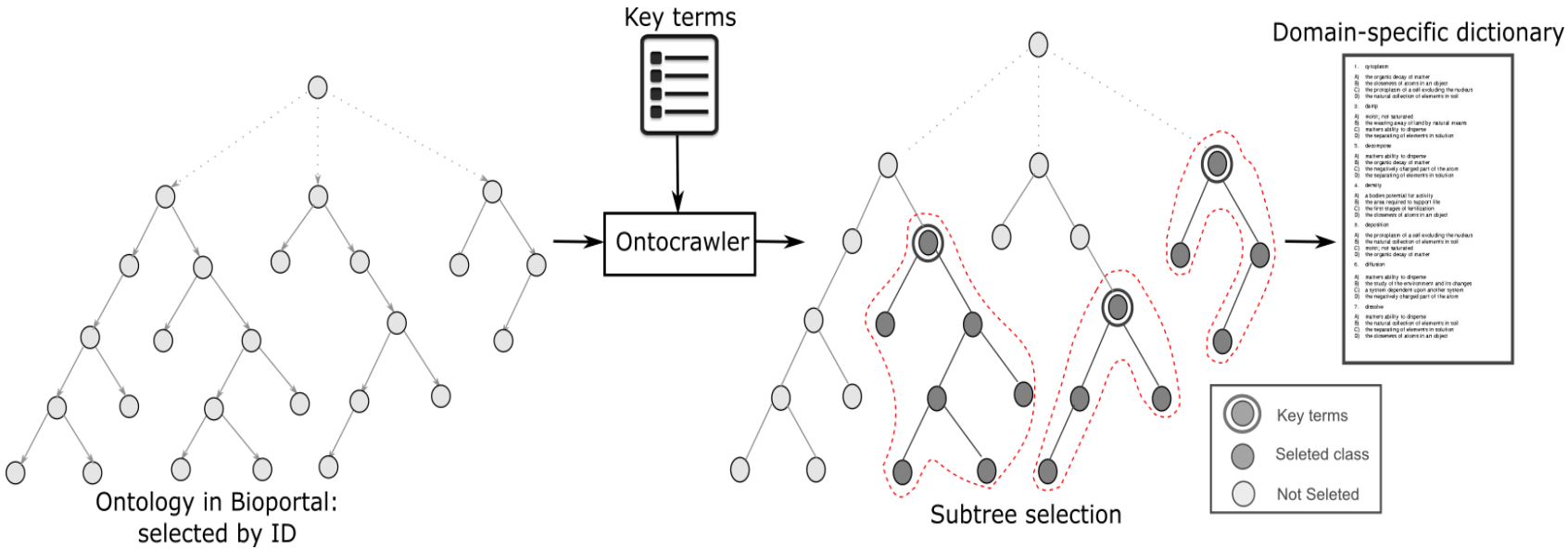
Input: Corpus of radiology reports



Output: Vector representation of reports

Ontocrawler: generation of domain dictionary

- Created an ontology crawler using SPARQL that grabs the sub-classes and synonyms of the domain-specific terms from NCBO bio-portal.
- Generate a focused dictionary for each domain of radiology.



- {‘apoplexy’, ‘contusion’, ‘hematoma’, ...} ➡ ‘hemorrhage’

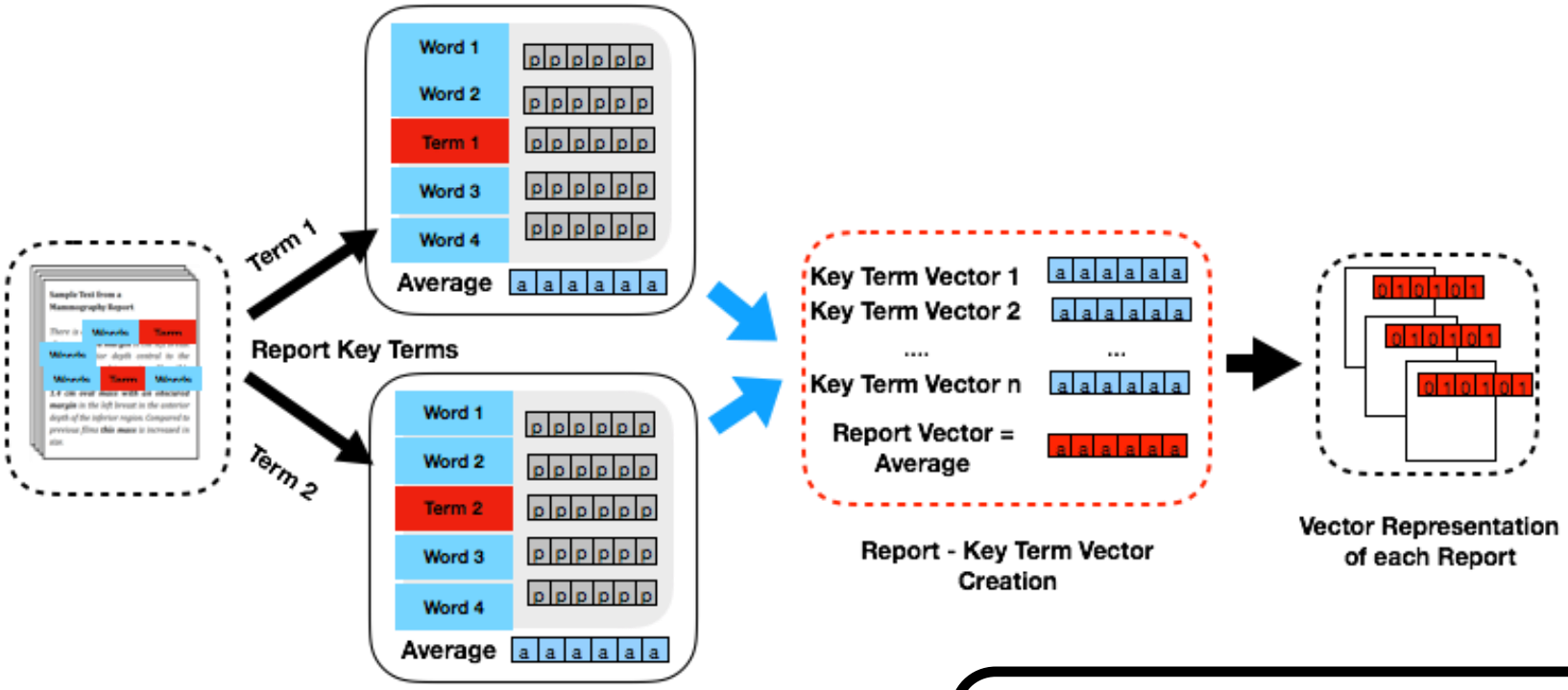
Stanford University

Context-depended document vector creation

Radiology impression: 1. Removal of right frontal subdural drain. Stable right subdural hematoma layering along the tentorium and posterior falx, and stable minimal leftward midline shift. 2. Paranasal sinus disease.

Term and context:

right	subdural	hemorrhage	layering	tentorium



For Document vector creation also used:

I. Averaging
II. Max pooling
III. Min pooling

Application of IWE

CT reports –

- Banerjee I, Madhavan S, Goldman RE, Rubin DL. Intelligent Word Embeddings of Free-Text Radiology Reports. AMIA Annual Symposium 2017
- Banerjee I, Chen MC, Lungren MP, Rubin DL. Radiology Report Annotation using Intelligent Word Embeddings. Journal of Biomedical Informatics November 2017
- Banerjee I. et. al., Comparative Effectiveness of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) Architectures for Radiology Text Report Classification, Journal of Artificial Intelligence in Medicine, 2018.

Mammograms –

- Imon Banerjee, Selen Bozkurt, Emel Alkim, Daniel L. Rubin, Automatic Inference of BIRADS Final Assessment Categories from Narrative Mammography Report Findings, Journal of Biomedical Informatics, (in press).

Ultrasound

- Imon Banerjee, Hailey H. Choi, Terry Desser, and Daniel L. Rubin. "A Scalable Machine Learning Approach for Inferring Probabilistic US-LI-RADS Categorization." , AMIA Annual Symposium (2018).
- 2 papers in RSNA 2018

Multiple clinical narratives

- Imon Banerjee, Kevin Li, …, James D. Brooks, Daniel L. Rubin, Tina Hernandez-Boussard, Weakly supervised natural language processing for assessing treatment-related side effects following prostate cancer treatment, JAMIA Open, 2019.
- Manuscript submitted to Journal of Clinical oncology

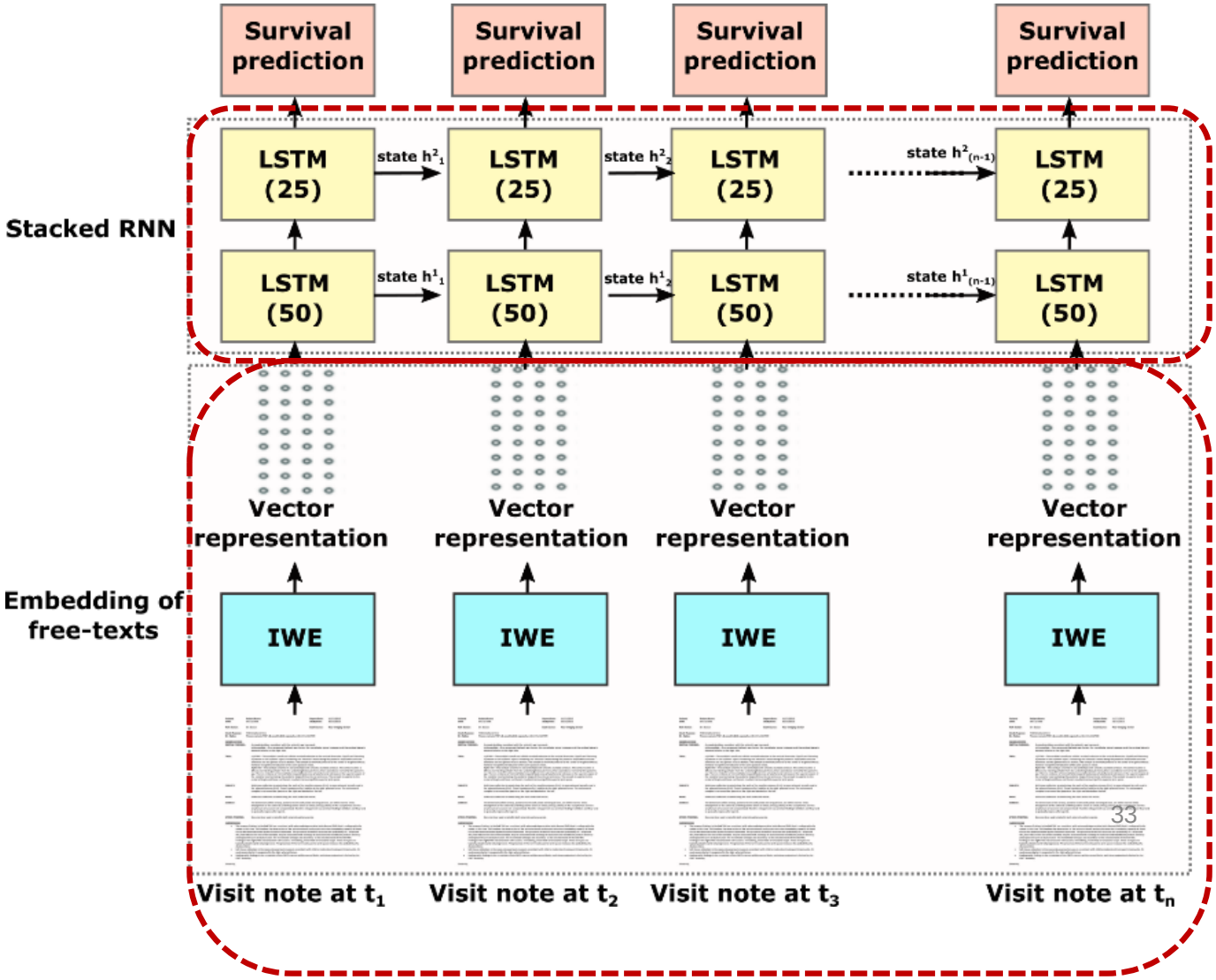
Study 1: Prognostic Estimates of Survival in Metastatic Cancer Patients (only notes)

- Only in United States around 500,000 patients develop metastatic cancer every year.
- Several studies have shown overutilization of aggressive medical interventions and protracted radiation treatment in patients close to the end of life.
- Inability to accurately estimate patient life expectancy likely explains why physicians tend to choose overly-aggressive treatments for some patients.
- Leads to increased morbidity and healthcare costs, while other patients may be under-treated and denied access to effective treatments that could reduce symptoms or even extend survival.

A robust ML model that predicts patient survival would have major impact on the quality of care and quality of life in metastatic cancer patients.

Banerjee I, Gensheimer MF, Wood DJ, Henry S, Chang D, Rubin DL. Probabilistic Prognostic Estimates of Survival in Metastatic Cancer Patients (PPES-Met) Utilizing Free-Text Clinical Narratives. Nature Scientific Reports

PPES-Met model



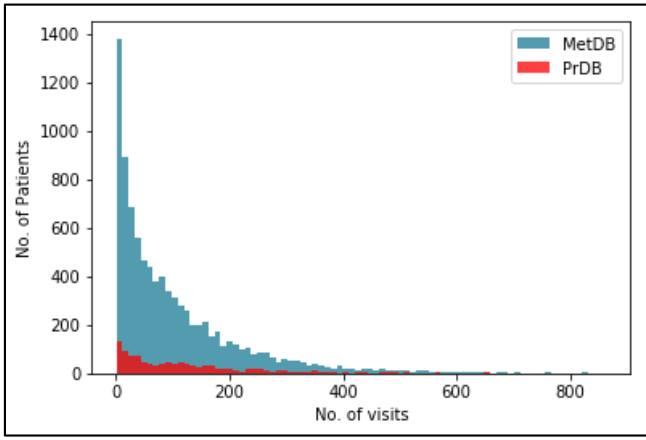
Many-2-many RNN model

Unsupervised embedding Of free-text notes

Stanford University

Dataset used in the study

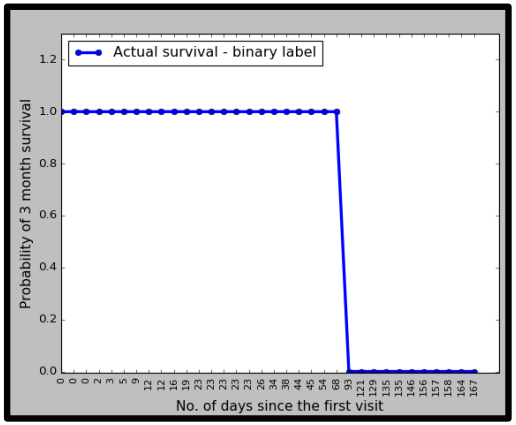
Characteristic	Metastatic cancer database (MetDB)	Palliative radiation dataset (PrDB)
No. of patients	13,523	899
Age	61.5 (IQR 51.2 – 70.5)	65.0 (IQR 55.8 – 72.2)
Sex	M: 6621 (49%); F: 6902 (51%)	M: 460 (51.1%); F: 439 (48.9%)
Primary site	Breast: 1493 (11.0%) Endocrine: 211 (1.6%) Gastrointestinal: 3575 (26.4%) Genitourinary: 1504 (11.1%) Gynecologic: 849 (6.3%) Head and neck: 506 (3.7%) Skin: 453 (3.3%) Thorax: 2178 (16.1%) Other/Multiple/Unknown: 2754 (20.4%)	Breast: 141 (15.7%) Endocrine: 0 (0%) Gastrointestinal: 145 (16.1%) Genitourinary: 112 (12.5%) Gynecologic: 50 (5.6%) Head and neck: 57 (6.3%) Skin: 122 (13.6%) Thorax: 252 (28.0%) Other/Multiple/Unknown: 20 (2.2%)
Note types	Oncology notes, progress notes, radiology reports, discharge summary, nursing notes, critical care notes	



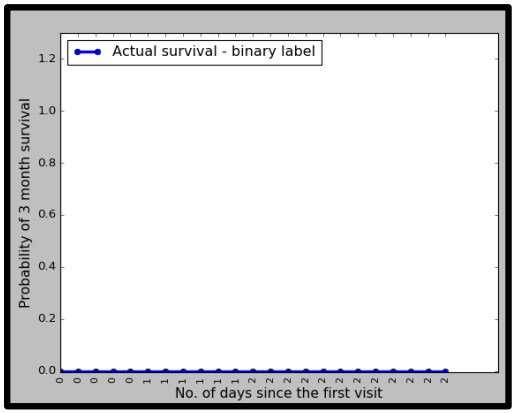
Distribution of visits

Stanford University

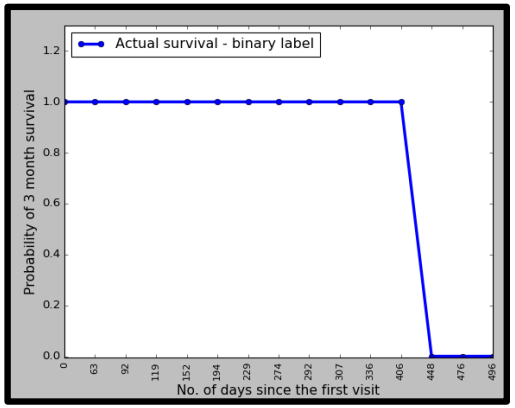
Survival data - challenges



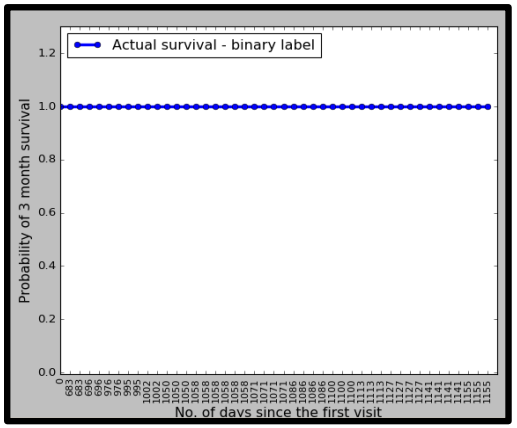
Patient 1: Dense follow up (multiple visits on same day)



Patient 2: Minimal information (only 3 days)



Patient 3: Sparse follow up (long and variable gaps between visits)

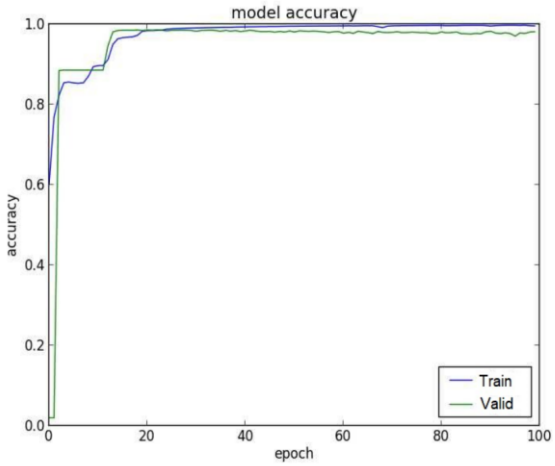


Patient 4: No death info – long follow up

Stanford University

Training and Evaluation

Model training and validation on MetDB



Training: 10,293 patients; Validation: 1,938 patients;

Category 1: "Survival - positive" stands for survival up to 3months starting from the current visit date;

Category 2: "Survival - negative" flagged the non-survival;

Category 3: "Zero padding" padded each input sequences when is shorter than 1000 and truncated the historic visits when sequence is longer than 1000

Model evaluation: dual strategy

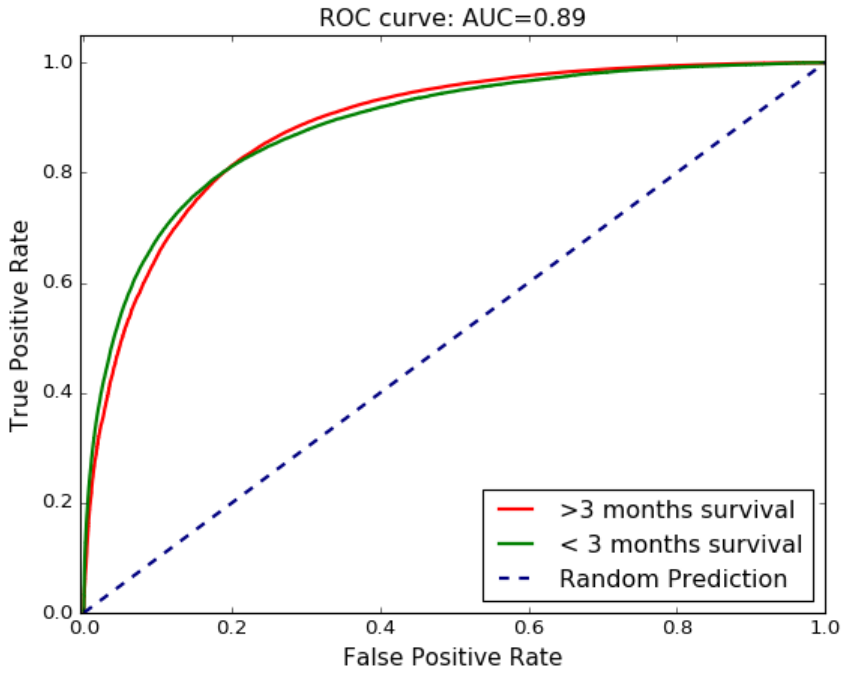
1. Quantitative: measure the overall prognosis estimation accuracy using the standard statistical metrics
2. Qualitative: evaluate the patient-level performance and perform error analysis with intelligible longitudinal graph summary for understanding the basis of prediction.

Test: 1818 patients; 899 from PrDB + 919 Randomly selected from MetDB

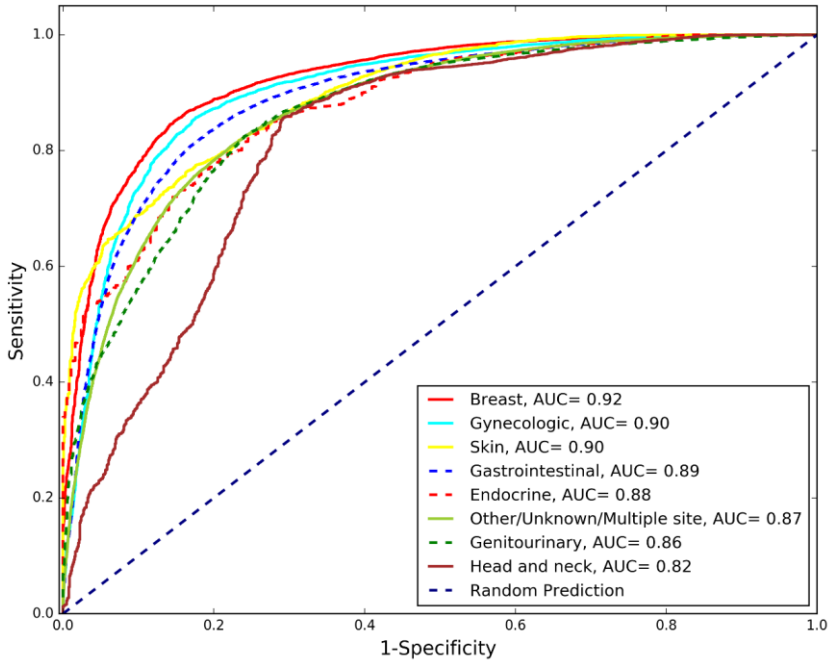
Stanford University

Results: Quantitative Evaluation on PrDB

Tested on 1818 patients with multiple visits



Overall ROCAUC for predicting 3 mo. survival - 0.89; Confidence interval [0.884 - 0.897]

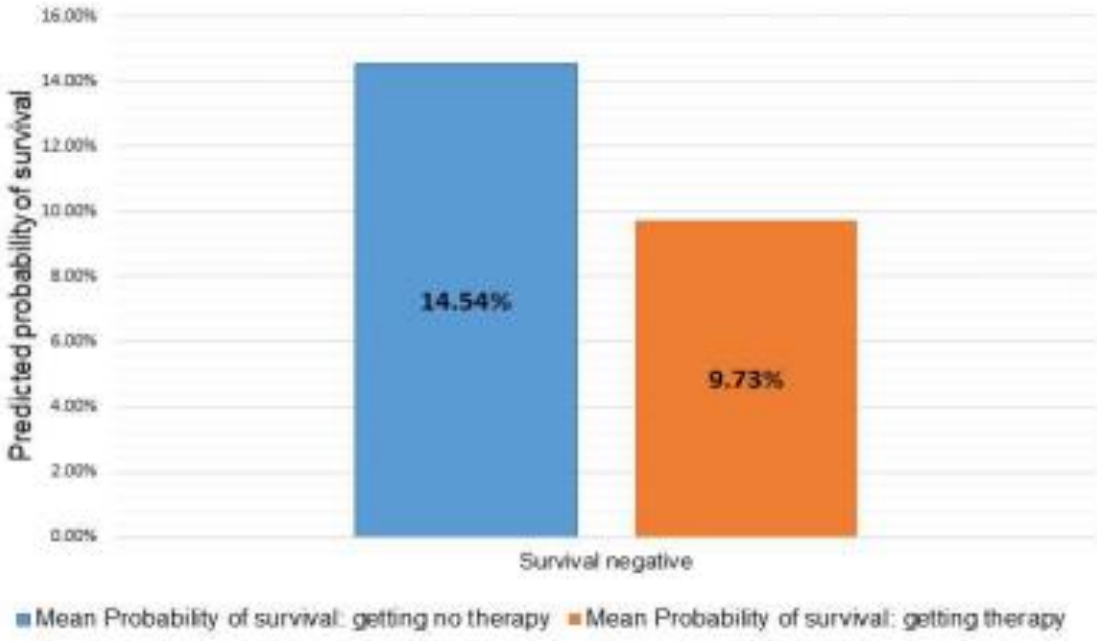


ROC based multiple primary site

Stanford University

Results: Quantitative Evaluation on PrDB

Tested on 1818 patients with different primary sites

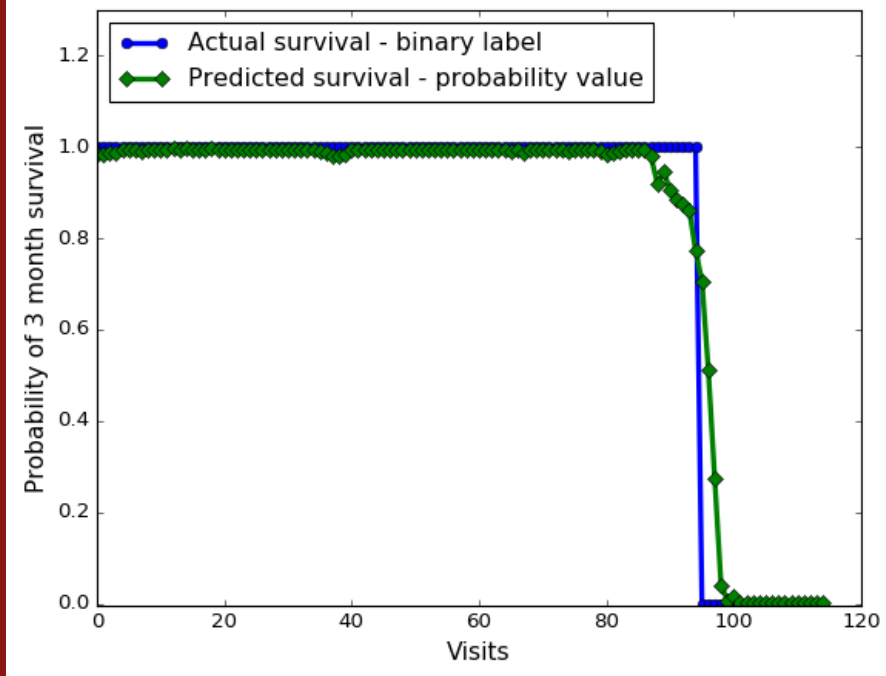


Comparing with systematic therapy:
Shows model's prediction outperformed oncologist's expectation of survival and can contribute in treatment planning

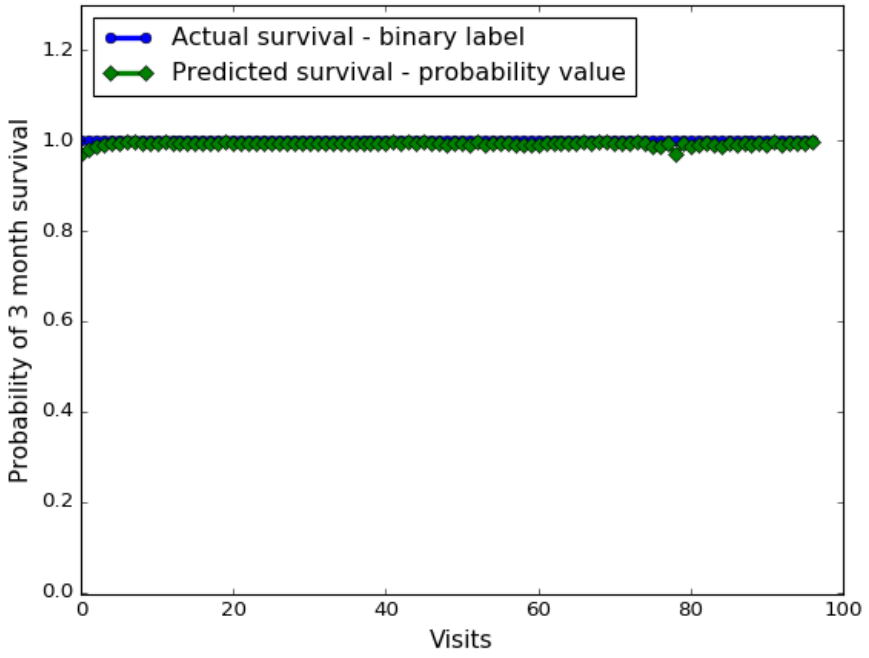
Stanford University

Results: Qualitative Evaluation on PrDB

Patient-level performance analysis



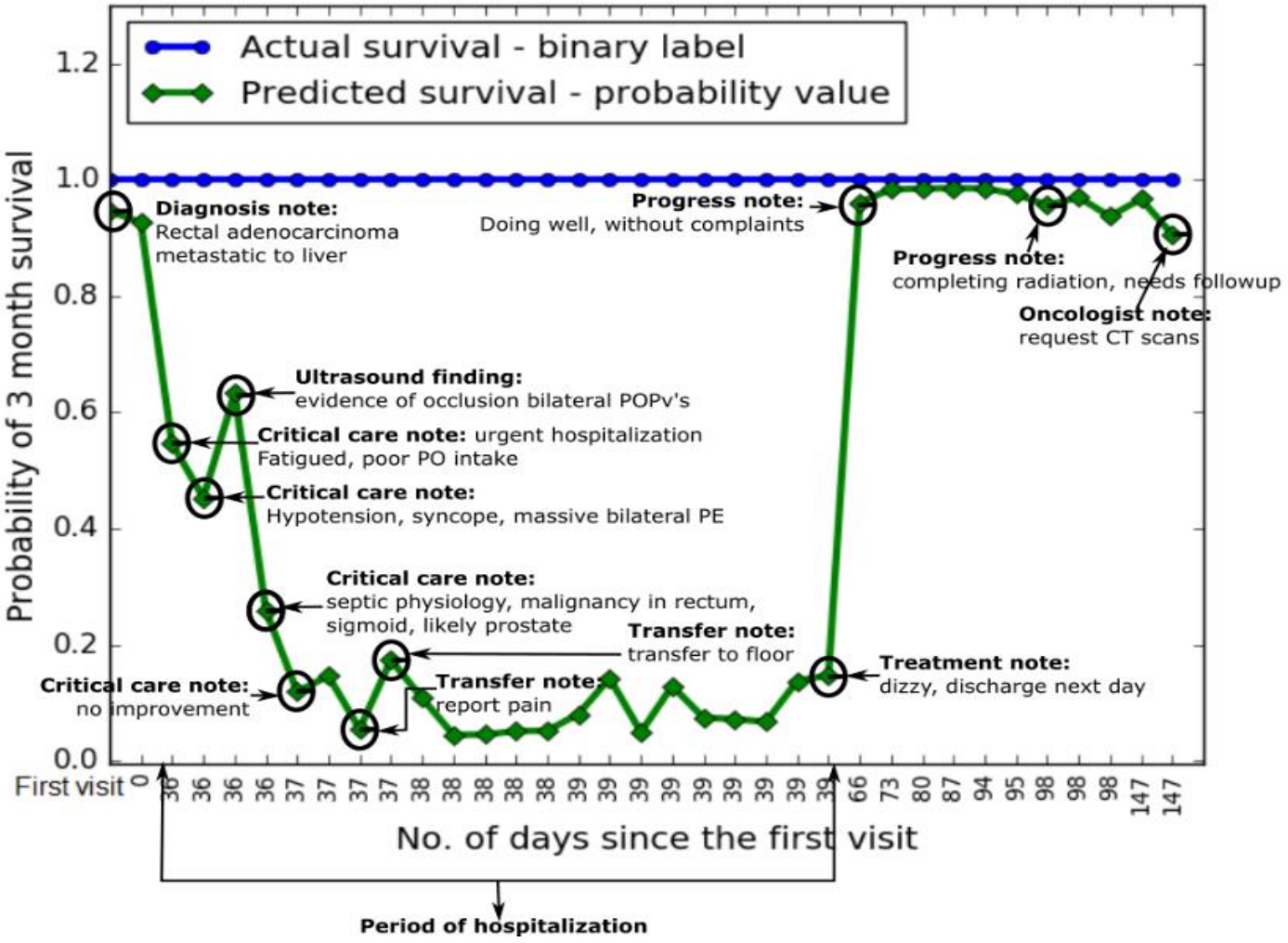
Patient 1



Patient 2

Stanford University

Hover & discover



Intelligible longitudinal survival curve of a patient

Stanford University

Study 2. Prognosis of AMD Disease using SD-OCT Imaging Biomarkers (Image + demograhics)

- Age-related macular degeneration (AMD) is the leading cause of visual loss
- Prediction of AMD progression may allow potential earlier treatment and better clinical outcomes.
- Most recent machine learning studies utilized genetic information and predicted the risk of AMD with high accuracy
- However, studied mainly in populations of European ancestry and predicted long-term AMD progression (>5-years).
- Image-based prediction models also showed success, but limited by mostly not considering dependencies of longitudinal visit data.

https://arxiv.org/abs/1902.10700

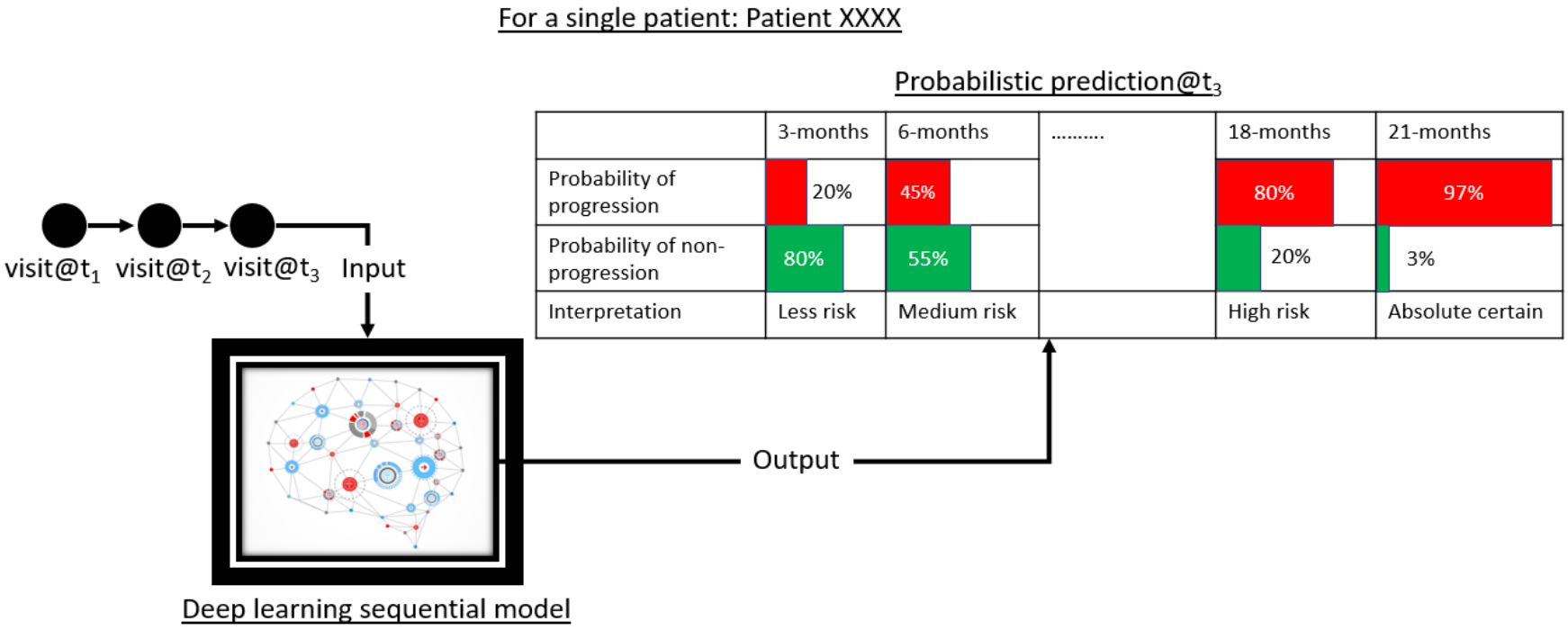
Stanford University

Objective

- Develop a sequential deep learning technique that can consider longitudinal visit data – SD-OCT images features and demographics
- Predict AMD progression using varying number of visit data with irregular time interval

- Short-term prediction: 3-months, 6-months, 9-months
- Long-term prediction: 12-months, 15-months, 18-months, 21-months

Conceptual model



	3-months	6-months	.....	18-months	21-months
Probability of progression	20%	45%		80%	97%
Probability of non-progression	80%	55%		20%	3%
Interpretation	Less risk	Medium risk		High risk	Absolute certain

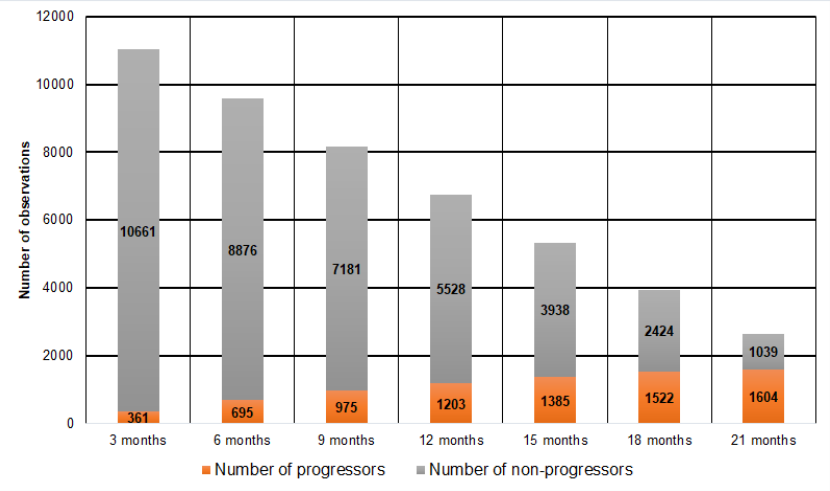
Stanford University

Dataset

HARBOR trial (ClinicalTrials.gov identifier: NCT00891735)
Patients had monthly evaluations with SD-OCT

Demographic Feature	Description	All fellow eyes (N=671)	Progressors (N=149)	Non-progressors (N=522)
Age	Age of the patient in months at baseline mean (std)	78.2 (8.3)	79.5 (7.7)	77.8 (8.4)
Gender	Patient gender: Male/Female %	40.4 / 59.6%	30.2 / 69.8%	43.3 / 56.7%
Race	Patient Ethnicity: American or Alaska native / Asian / Black or African American / White / Native Hawaiian or Pacific Islander / Multiracial	0.3 / 1.6 / 0.4 / 96.9 / 0.3 / 0 %	0 / 0.7 / 0 / 98.7 / 0.7 / 0 %	0.4 / 1.9 / 0.6 / 96.4 / 0.2 / 0 %
Smoking status	Smoking status: Non-smoker / Previous smoker / Current smoker	41.0 / 48.4 / 10.6 %	38.9 / 47.0 / 14.1 %	41.6 / 48.8 / 9.6 %
Visual Acuity	Visual acuity at baseline of observation measured in LogMAR scale	76.07 (13.07)	76.91 (9.31)	75.83 (13.96)

Demographic considered in our analysis

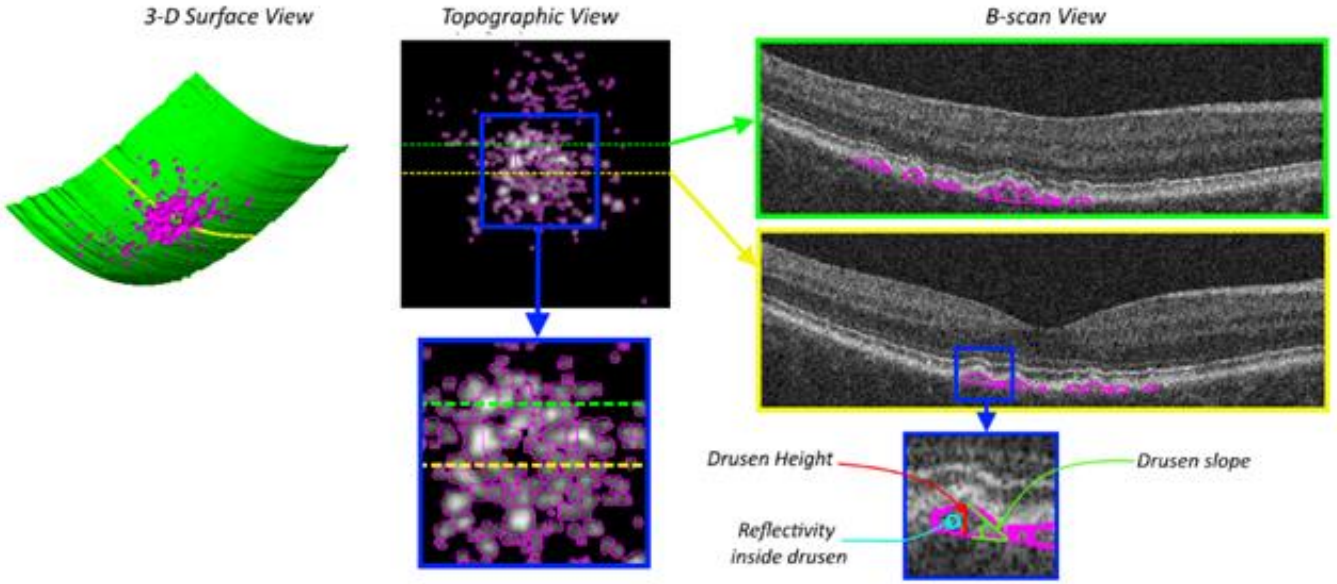


Counts of observations

Stanford University

Extraction of imaging biomarkers

- Each OCT volume was processed using proprietary Cirrus Review Software and a previously published pipeline[1]
- 21 imaging features describing presence, number, extent, density and relative reflectivity of drusen were extracted



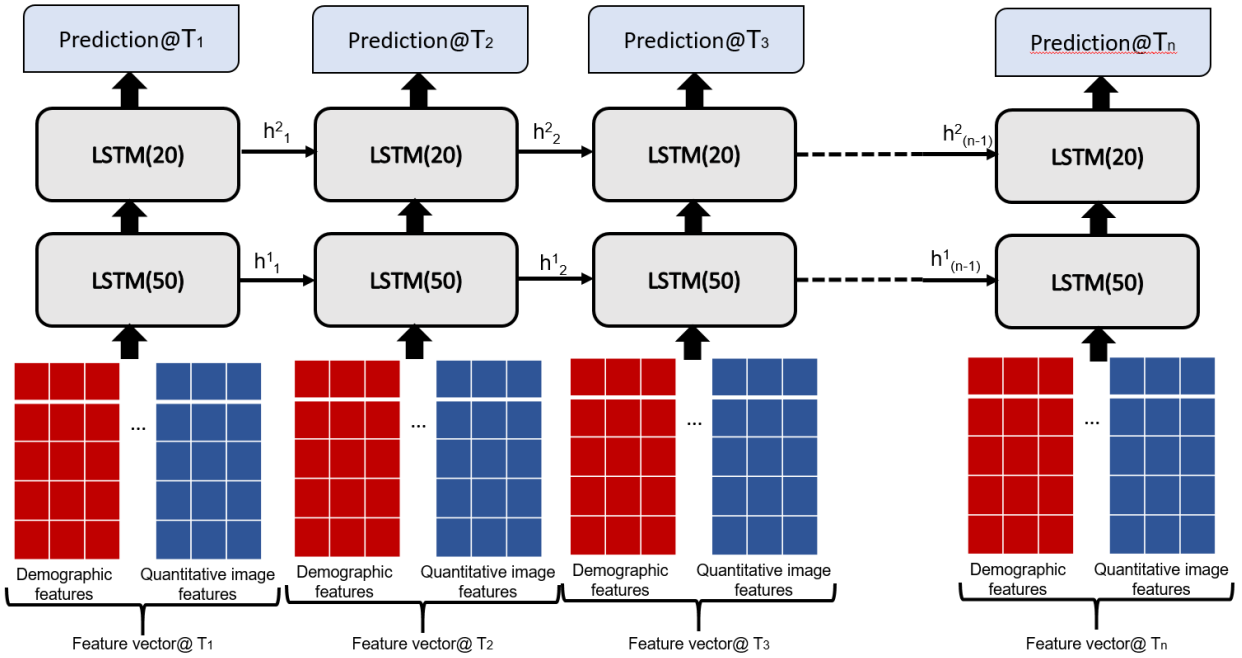
OCT image processing pipeline - overview

[1]de Sisternes, Luis, et al. "Quantitative SD-OCT imaging biomarkers as indicators of age-related macular degeneration progression." *Investigative ophthalmology & visual science* 55.11 (2014): 7093-7103.

Stanford University

Deep sequential model - RNN

- Designed a many-to-many RNN model using two-layer one-directional stacked stateful Long short-term memory (LSTM)
- Long-term memory allows slow weight updates during training and encodes general information about the whole temporal visit sequence
- Short-term memory has ephemeral activation and passes immediate state between successive nodes for resetting itself if a fatal condition is encountered.



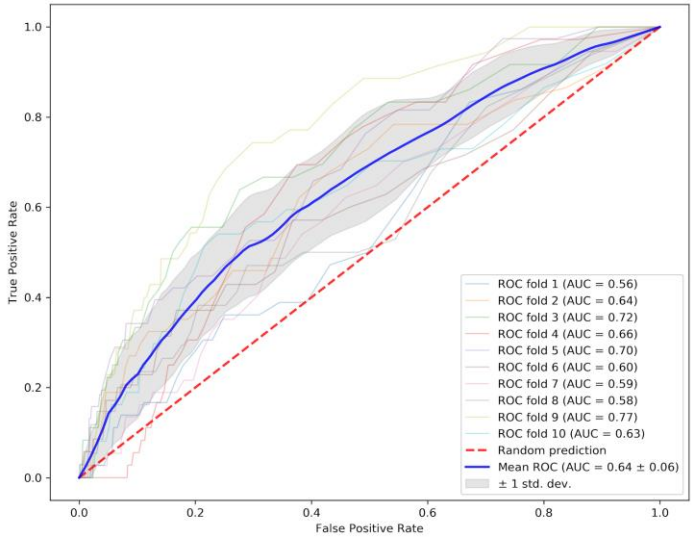
Deep sequential model: architecture

Stanford University

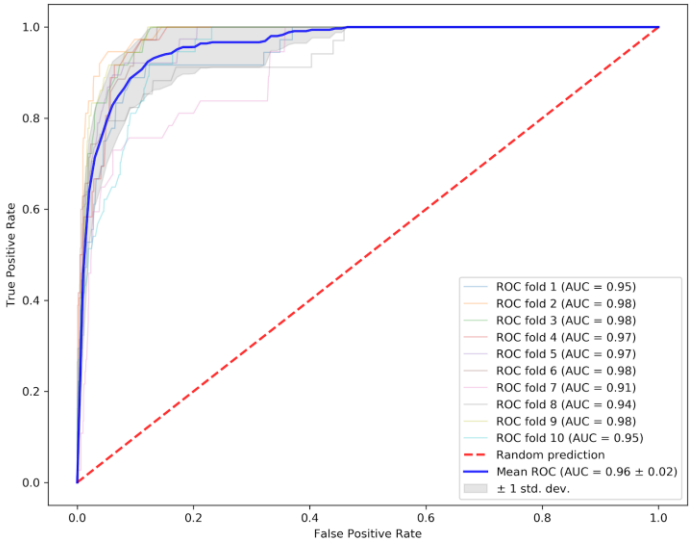
Short-term prediction – Comparison

Evaluated on a 10-fold cross validation setting where the original sample (13,954 time points of total 671 patients)

Random Forest: 0.64+/-0.06 AUC



Deep Sequence: 0.96+/-0.02



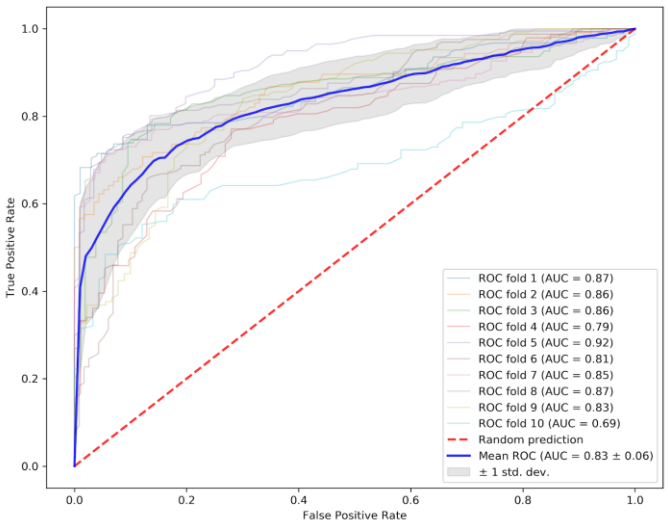
10-fold cross validation ROC curves: 3-months prediction

	AUC-ROC 10-fold Cross-validation	
	Random Forest	Deep Sequence
6-months	0.63+/-0.05	0.83+/-0.04
9-months	0.62+/-0.06	0.79+/-0.01

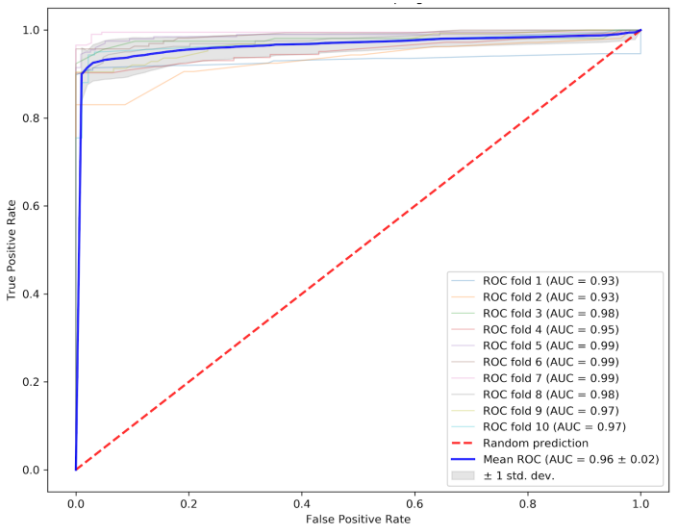
Stanford University

Long-term prediction – Comparison

Random Forest: 0.83+/-0.06 AUC



Deep Sequence: 0.96+/-0.02

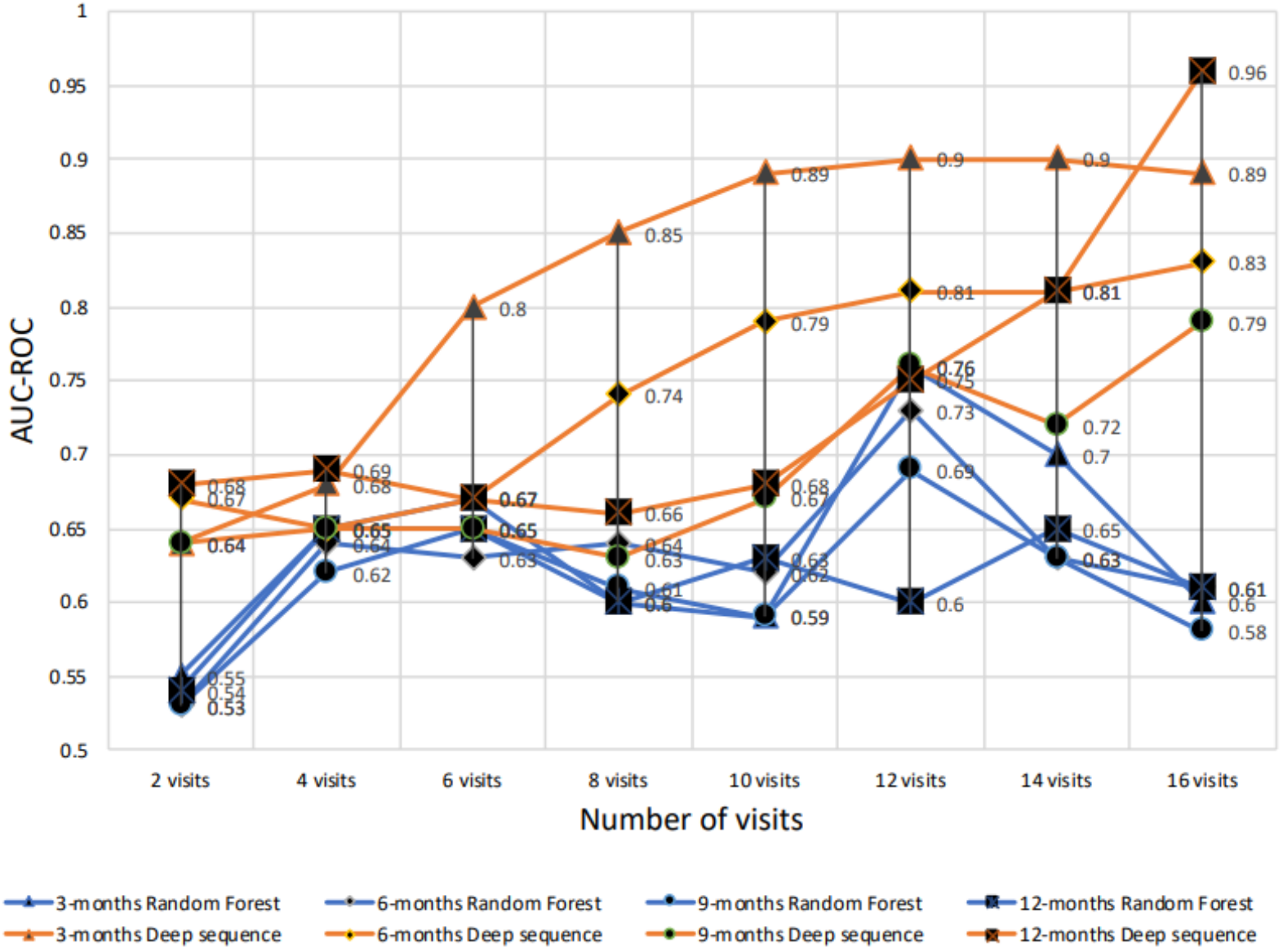


10-fold cross validation ROC curves: 21-months prediction

	AUC-ROC 10-fold Cross-validation	
	Random Forest	Deep Sequence
12-months	0.64+/-0.06	0.77+/-0.06
15-months	0.69+/-0.06	0.84+/-0.08
18-months	0.74+/-0.06	0.89+/-0.05

Stanford University

Prediction with varying number of visits in sequence



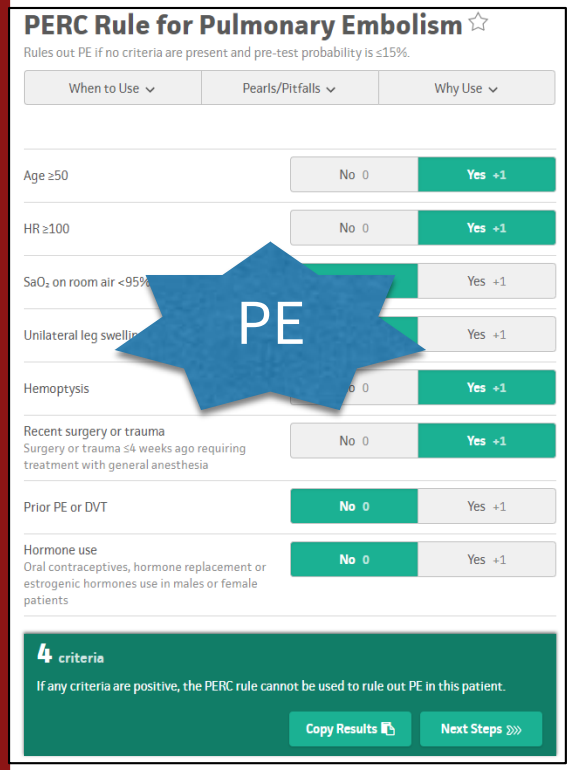
Study 3. Risk score assessment for PE (Structured EMR)

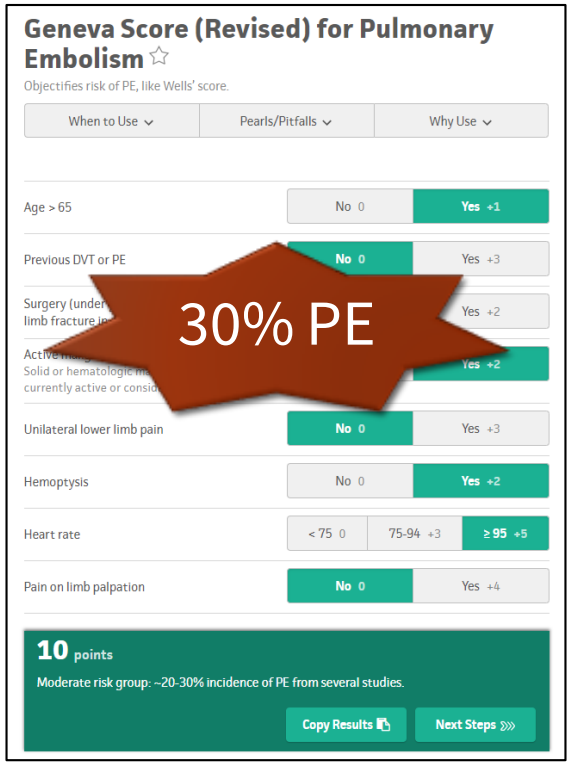
- 27-fold increase in the total number of CT angiography examinations performed for PE evaluation
- Rate of positive studies declined from 27% to less than 10%
- It has been reported that up to one-third of all PE-CTA imaging studies are avoidable
- Many problems exist with current guideline and contributes to clinician noncompliance.
- Other known clinical risk factors NOT included in ANY risk scores:

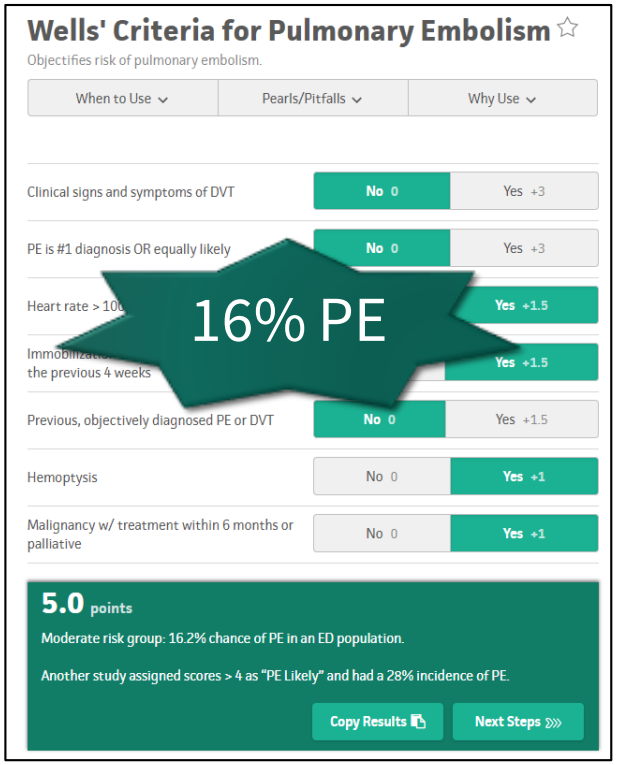
Currently in preparation

Stanford University

PE clinical scorings

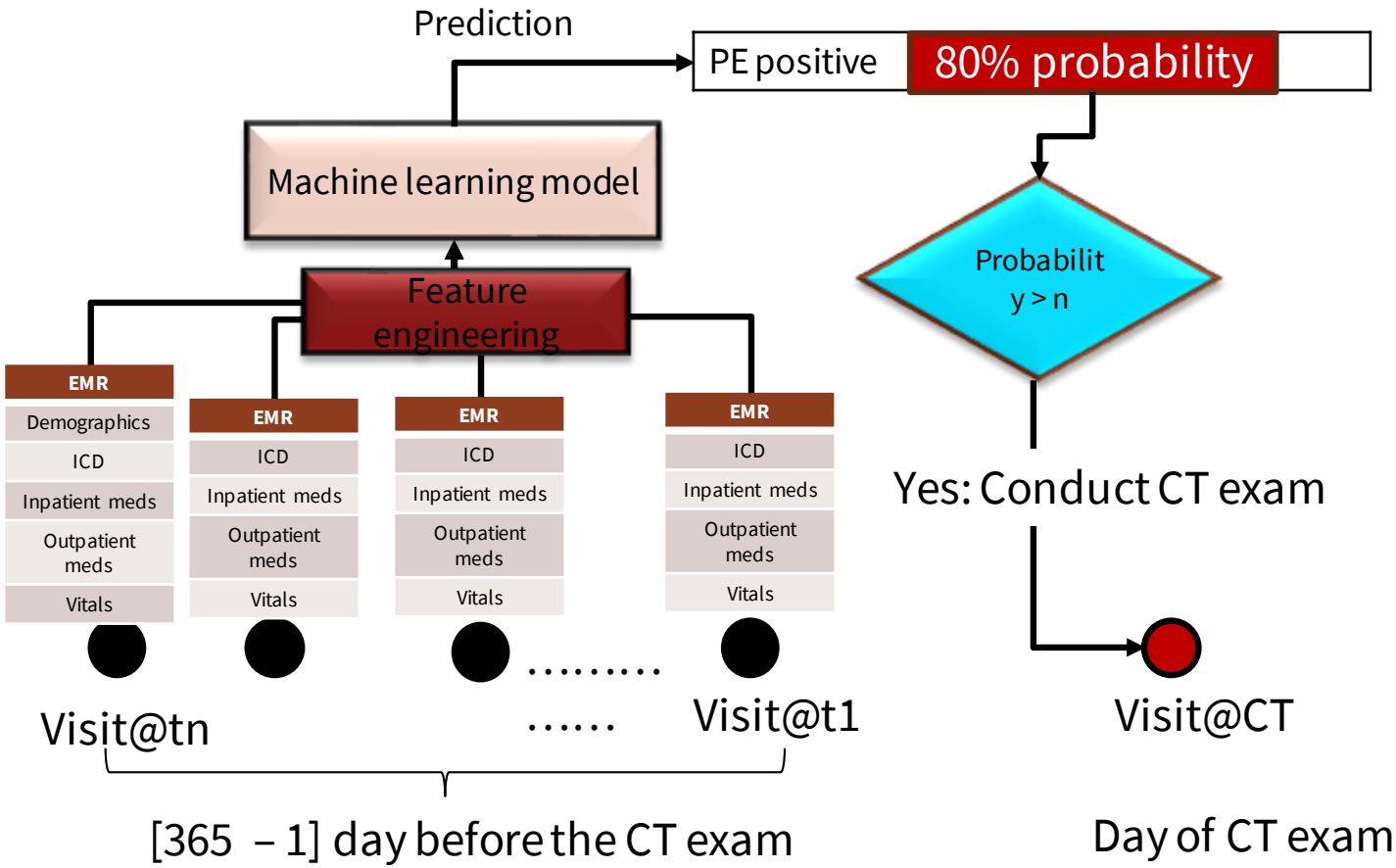






Stanford University

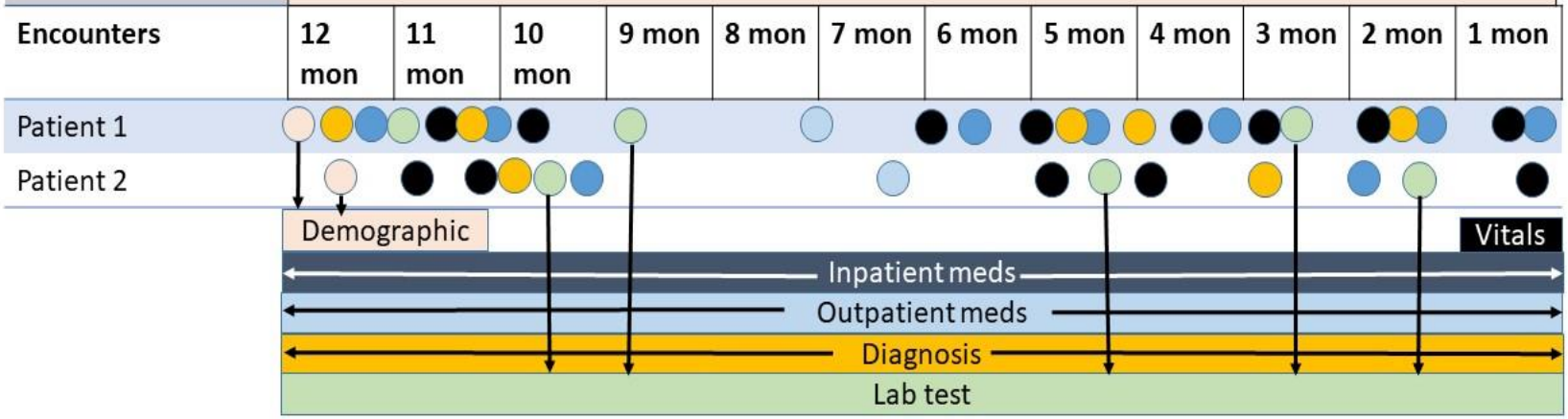
Prediction model – Structured EHR only



Stanford University

Temporal feature engineering

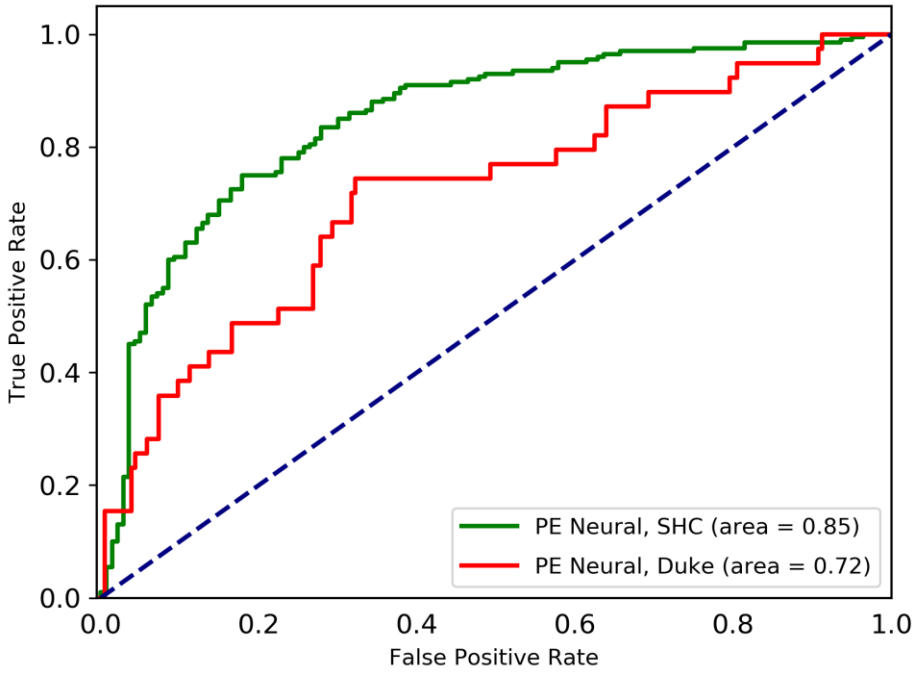
Information considered	Prior encounters difference with CT date											
	12 mon	11 mon	10 mon	9 mon	8 mon	7 mon	6 mon	5 mon	4 mon	3 mon	2 mon	1 mon
Vitals derivation												$\partial x/\partial t$
Inpatient Meds	614 unique Pharma class: Presence and Frequency of order											
Outpatient Meds	614 unique Pharma class: Presence and Frequency of order											
Diagnosis code	141 unique code: Presence of code											
Laboratory tests	21 unique test: Presence of test and latest value											
Demographics	Age, Gender, Race, Smoking habit: Only latest value											



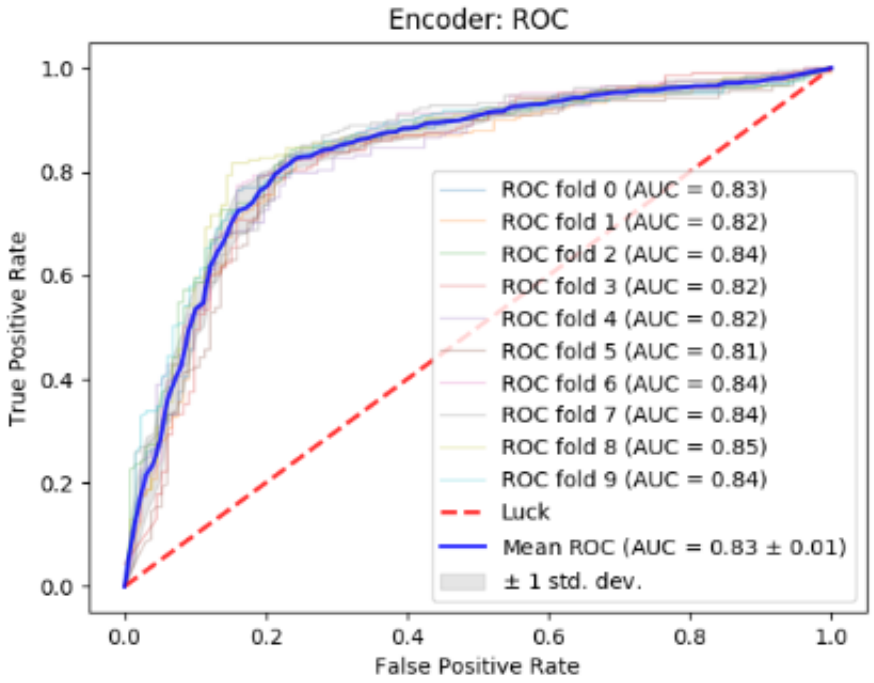
Stanford University

Prediction with autoencoder with attention

Hold-out testset: Stanford 340 cases, Duke 244 cases

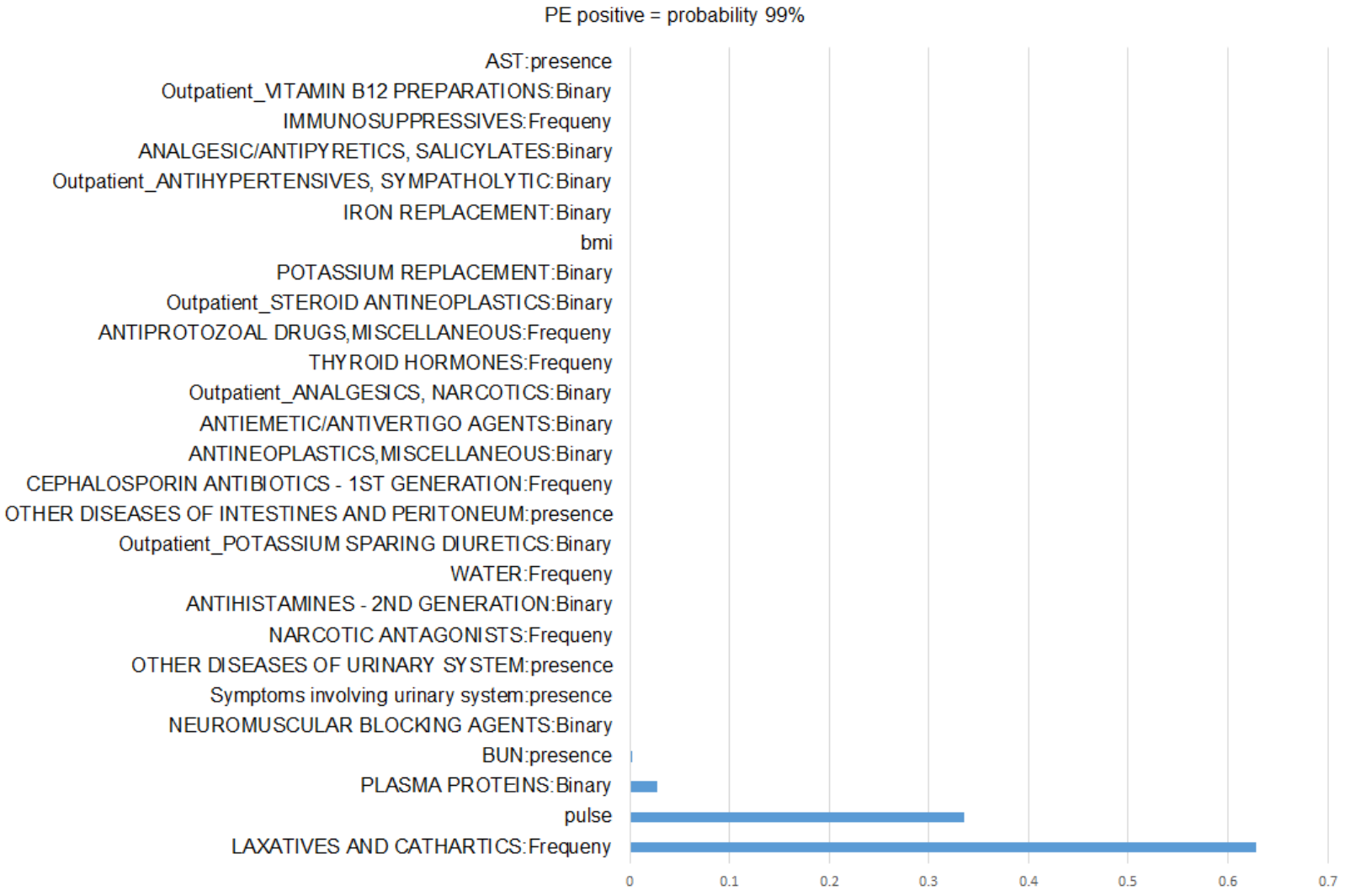


10-fold cross-validation



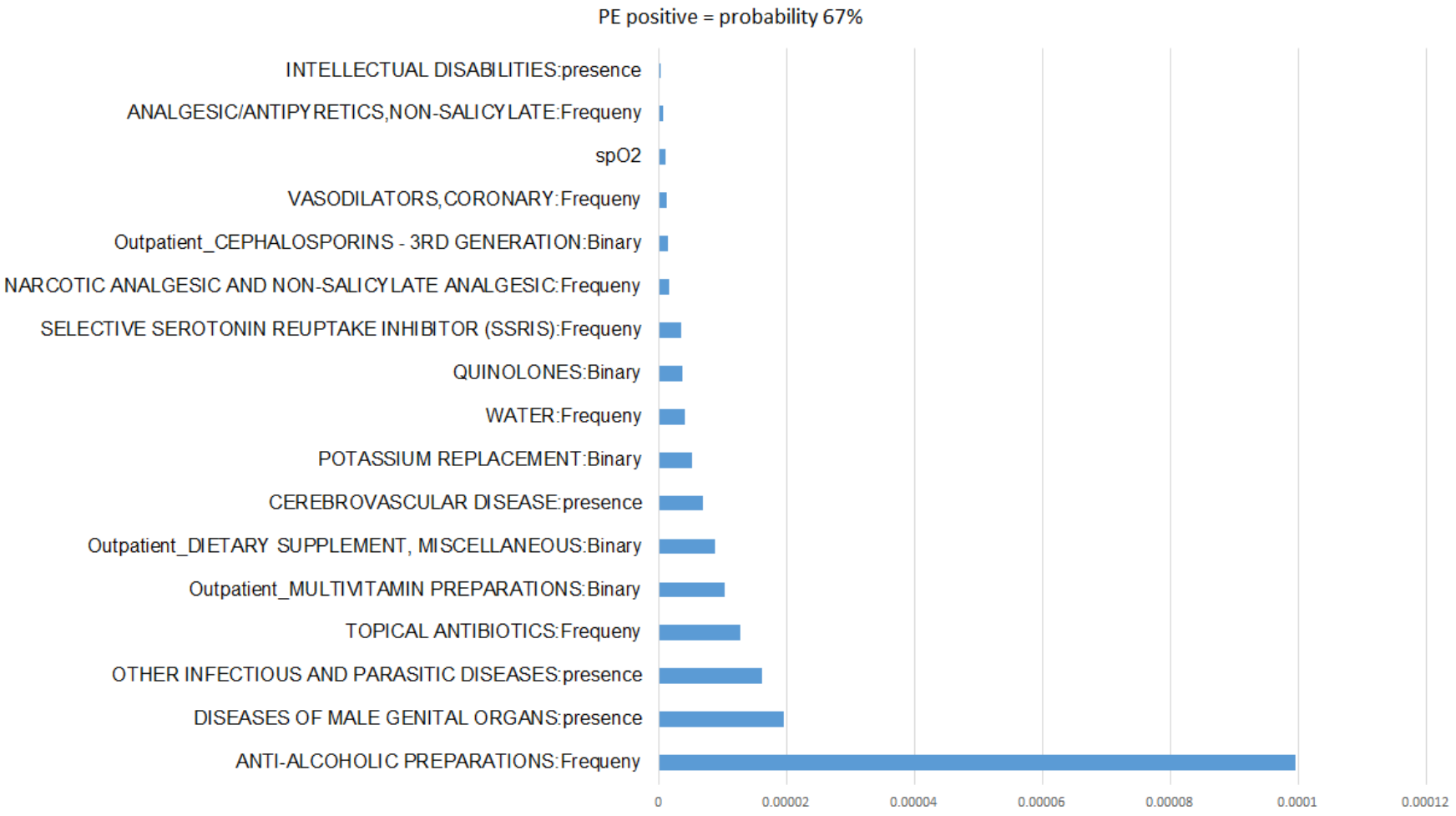
Stanford University

Case 1



Stanford University

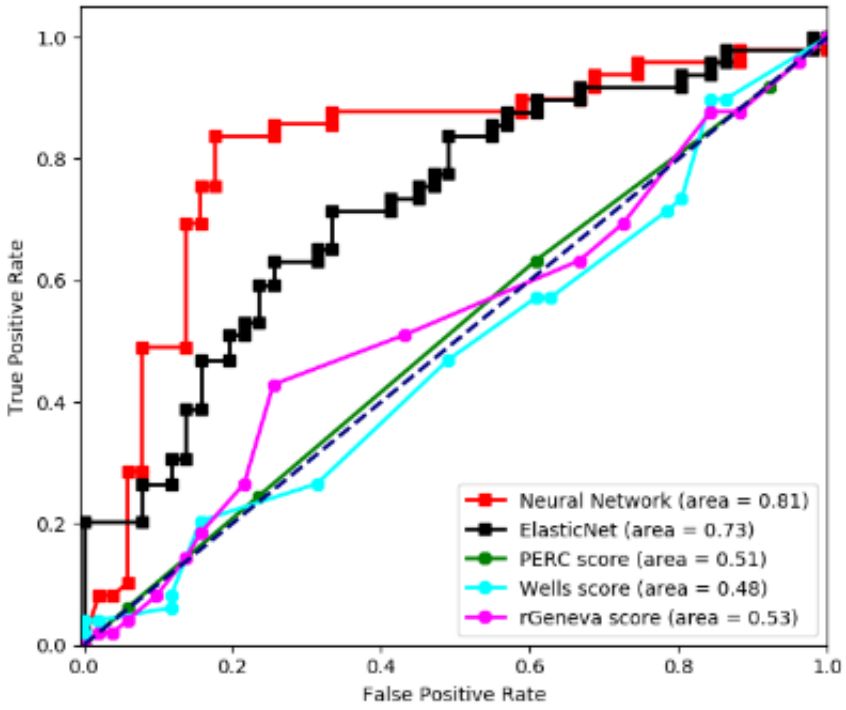
Case 2



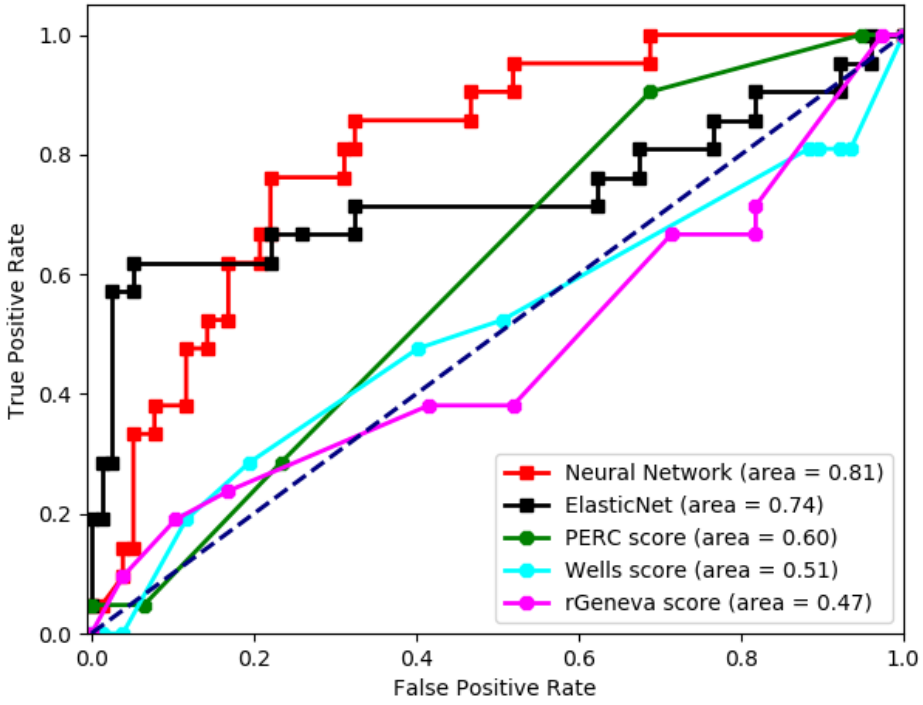
Stanford University

Comparison with clinical scoring

100 Stanford ED patients - manual chart review



100 Duke ED patients - manual chart review



Stanford University

Outline

1. Need for image interpretation beyond image classification
2. Integrating multiple data types with images
3. Making AI clinical predictions and providing visualizations for explanation
4. Evaluation of AI algorithms

Importance of evaluating AI systems

Everything an AI system "knows" is based on the **data upon which it is trained**

AI algorithms may not **generalize** to new data (wasn't seen before)

- Data used to create algorithms can contain **bias**
- Differences in **patient populations** (e.g., foreign vs. domestic
- Differences in **equipment/parameters** for imaging
- Rare **disorders/abnormalities** may be under-represented

Example: Pneumonia detection

158,323 chest radiographs from three institutions

- NIH (30,805 patients)
- Mount Sinai Hospital (MSH; 12,904 patients)
- Indiana (IU; 3,807 radiographs from 3,683 patients)

Task: Detecting radiographic findings consistent with pneumonia

Result: AI trained on data from **individual** or **multiple hospital systems** did not consistently ***generalize*** to external sites

Zech JR et al., Confounding variables can degrade generalization performance of radiological deep learning models, arXiv:1807.00431

Pneumonia detection (cont'd)

Train - Tune Site	Comparison Type*	Test Site (Images)	AUC (95% C.I.)	Acc.	Sens.	Spec.	PPV	NPV
NIH	Internal	NIH (N=22,062)	0.750 (0.721-0.778)	0.255	0.951	0.247	0.015	0.998
	External	MSH (N=8,388)	0.695 (0.683-0.706)	0.476	0.950	0.212	0.401	0.884
	External	IU (N=3,807)	0.725 (0.644-0.807)	0.190	0.974	0.182	0.012	0.999
	Superset	MSH + NIH (N=30,450)	0.773 (0.766-0.780)	0.462	0.950	0.403	0.160	0.985
	Superset	MSH + NIH + IU (N=34,257)	0.787 (0.780-0.793)	0.470	0.950	0.418	0.148	0.987
MSH	Internal	MSH (N=8,388)	0.802 (0.793-0.812)	0.617	0.950	0.432	0.482	0.940
	External	NIH (N=22,062)	0.717 (0.687-0.746)	0.184	0.951	0.175	0.014	0.997
	External	IU (N=3,807)	0.756 (0.674-0.838)	0.099	0.974	0.090	0.011	0.997
	Superset	MSH + NIH (N=30,450)	0.862 (0.856-0.868)	0.562	0.950	0.516	0.190	0.989
	Superset	MSH + NIH + IU (N=34,257)	0.871 (0.865-0.877)	0.577	0.950	0.537	0.180	0.990
MSH + NIH	Internal	MSH + NIH (N=30,450)	0.931 (0.927-0.936)	0.732	0.950	0.706	0.279	0.992
	Subset	NIH (N=22,062)	0.733 (0.703-0.762)	0.243	0.951	0.234	0.015	0.997
	Subset	MSH (N=8,388)	0.805 (0.796-0.814)	0.630	0.950	0.451	0.491	0.942
	External	IU (N=3,807)	0.815 (0.745-0.885)	0.238	0.974	0.230	0.013	0.999
	Superset	MSH + NIH + IU (N=34,257)	0.934 (0.929-0.938)	0.732	0.950	0.709	0.258	0.993

*Superset= a test dataset containing data from the same distribution (hospital system) as the training data as well as external data. Subset = a test dataset containing data from fewer distributions (hospital systems) then the training data.

Zech JR et al., Confounding variables can degrade generalization performance of radiological deep learning models, arXiv:1807.00431

Stanford University

Conclusion:

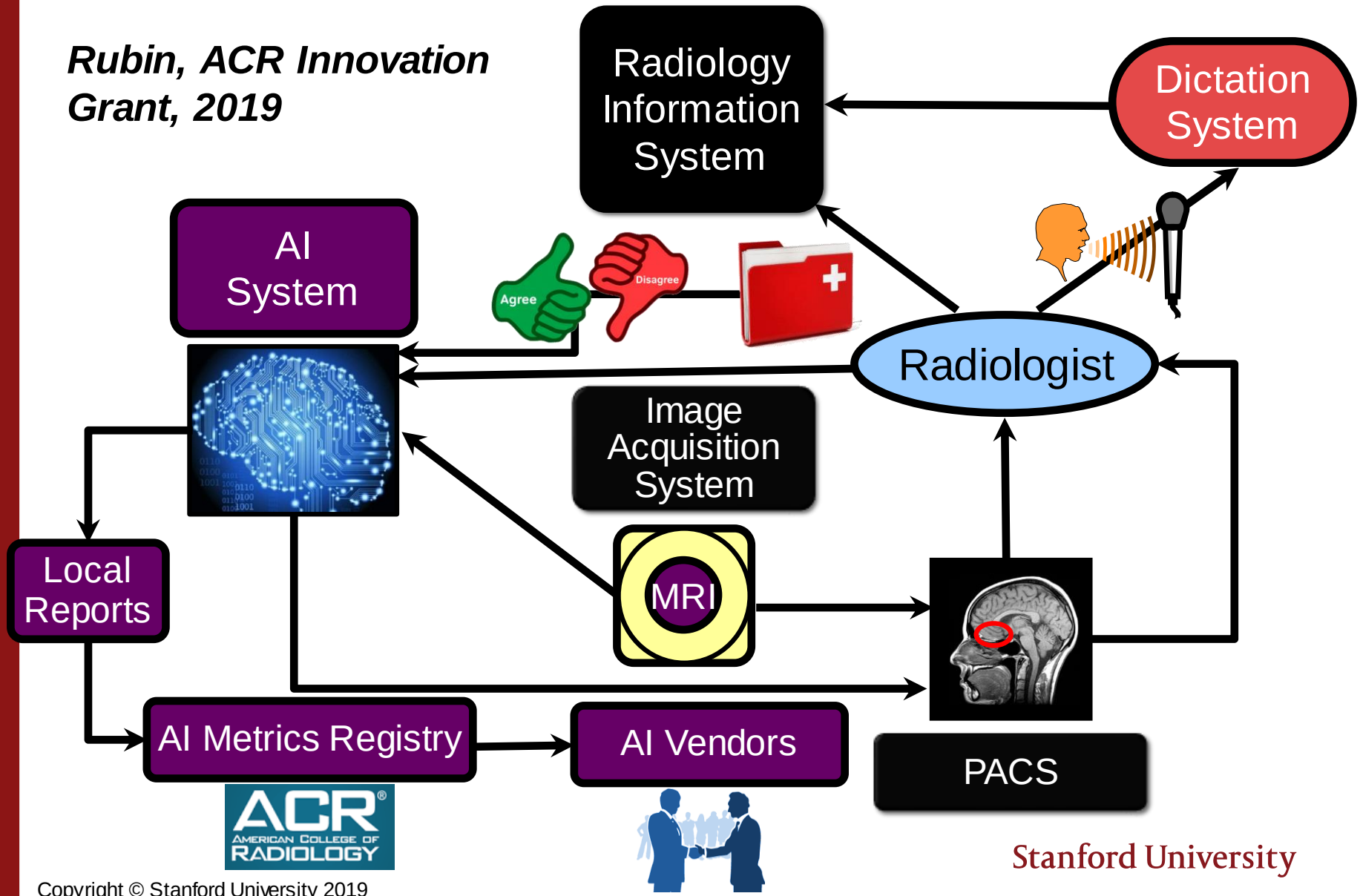
Need to test (and monitor) performance of AI on real-world data as part of adoption in clinical practice

Steps for undertaking evaluation

1. Understand the key outputs of the AI algorithm (what is it predicting?) and decide if that is clinically relevant to your needs
2. Choose appropriate evaluation metric (e.g., sensitivity, specificity, PPV)
3. Define performance threshold for the metric (e.g., 99% sensitivity in detecting cancer; this sets a **threshold on false positives**)
4. Collect representative patient samples (test cases)
5. Establish **ground truth** for each test case
6. Evaluate the test cases against the metric
7. (Implement monitoring strategy)

Collecting AI performance metrics in the clinical workflow

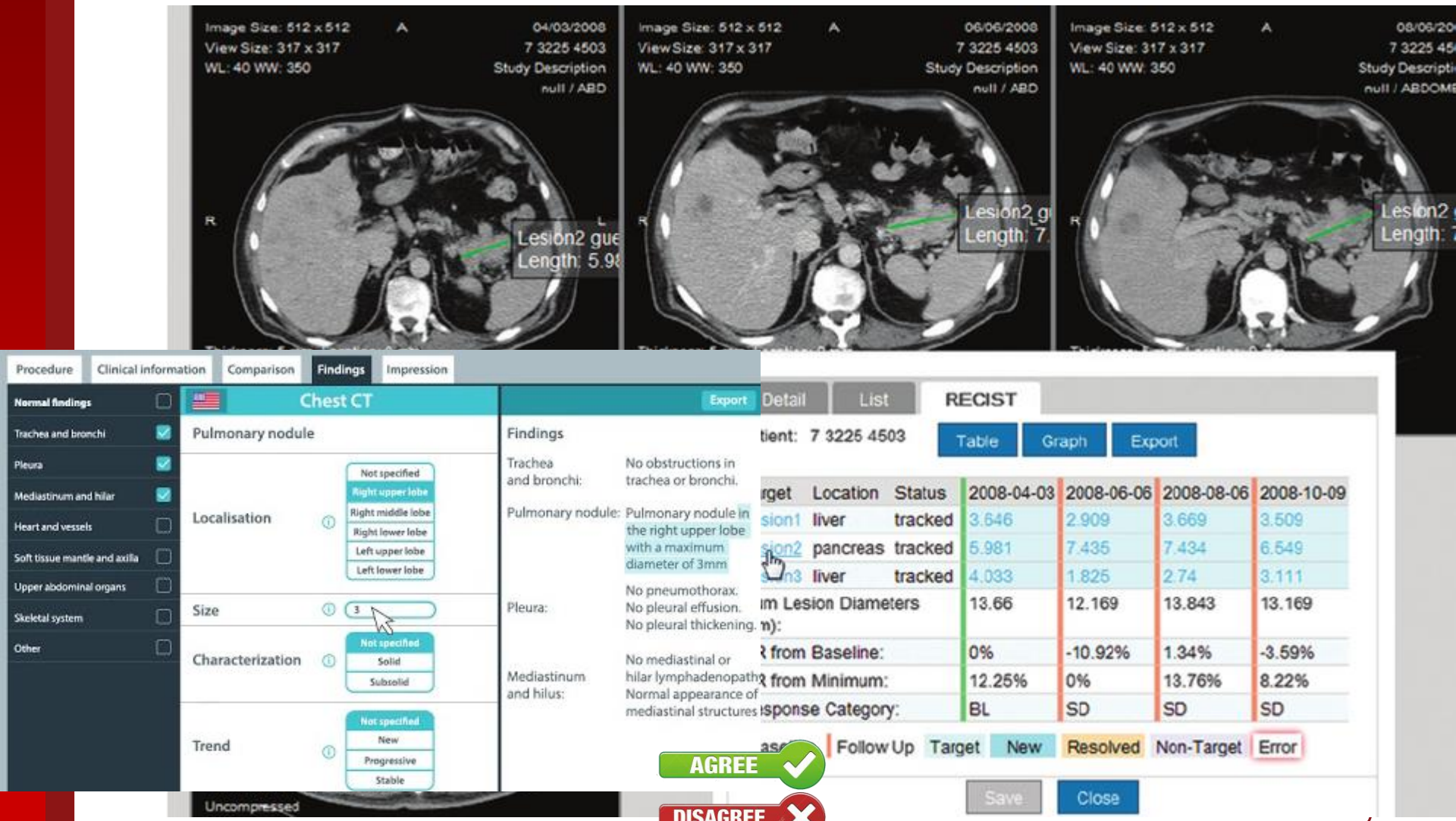
Rubin, ACR Innovation Grant, 2019



Stanford University

Toolkit for collecting AI performance metrics in the clinical workflow

Rubin, ACR Innovation Grant, 2019



Conclusion

- There are opportunities and needs to develop AI algorithms for medical problems other than **classification problems**, especially prediction
- Tackling **clinical prediction** enabled by integrating multiple data
  - Images, texts, and other data
  - Longitudinal time points
- **Preliminary work** applying AI to clinical predication problems is promising
- **Evaluation of AI algorithms** developed in actual clinical practice is crucial

Thank you.

Contact info:
dlrubin@stanford.edu
imonb@stanford.edu