ARTIFICIAL INTELLIGENCE IN HEALTH CARE

RICHARD VAN ECK
ASSOCIATE DEAN FOR TEACHING AND LEARNING
MONSON ENDOWED CHAIR FOR MEDICAL EDUCATION

PRESENTATION FOR THE DAKOTA CONFERENCE ON RURAL AND PUBLIC HEALTH
JUNE 4–6

1

DISCLOSURES

- No financial disclosures
- Some of this information comes from an MIT online course, including examples and some figures and images. These are identified and credited.

2

OBJECTIVES

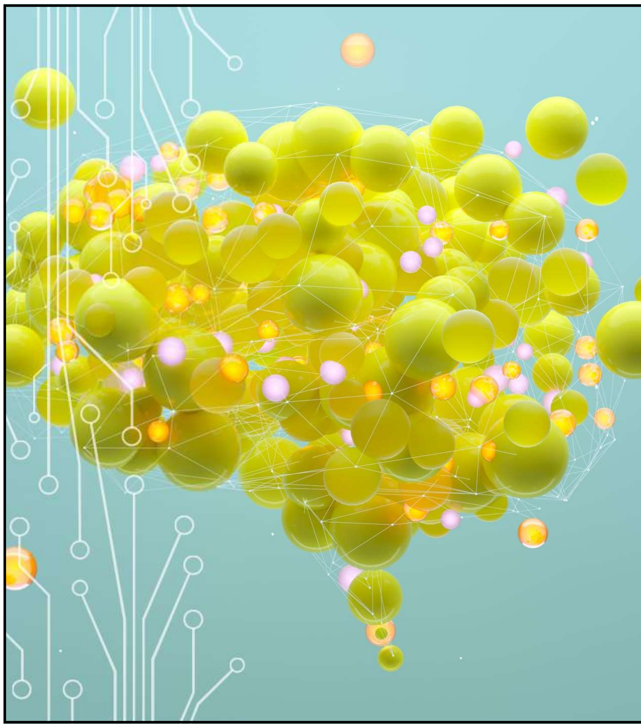
- Define supervised and unsupervised learning
- Define natural language processing, deep learning, neural networks
- Define the process by which supervised learning models are created, trained, and deployed
- List key considerations for designing deep learning models, including cost, bias, risk, explainability, and interpretability
- List examples of how AI is being used in health care environments.

3

OVERVIEW

- Defining AI
  - Supervised Machine Learning
    - Natural Language Processing
    - Neural Networks/Deep Learning
  - Unsupervised Machine Learning
    - Generative AI
- Customized AI Solutions
  - Training Data Sets
  - Annotation
  - Curation
- Risk, Bias, Best Practices
  - Interpretability/Explainability
  - Adversarial Training
- Case Studies of AI in Health Care
  - Hospital Optimization
  - Reducing Burden
  - Augmenting Clinical Workflows
  - Risk Stratification
  - Mental Health
  - Invisibles

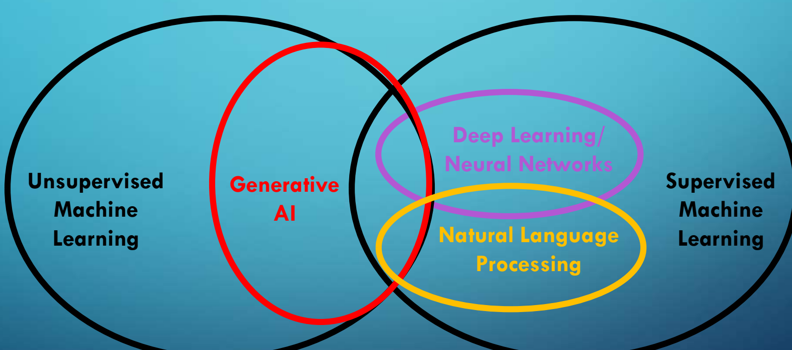
4



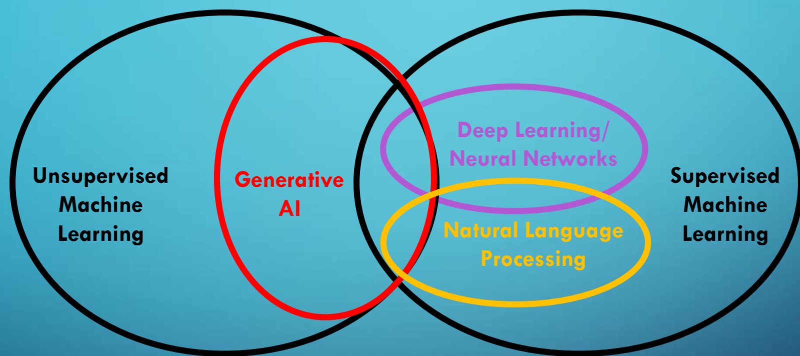
DEFINING AI

5

DEFINING ARTIFICIAL INTELLIGENCE

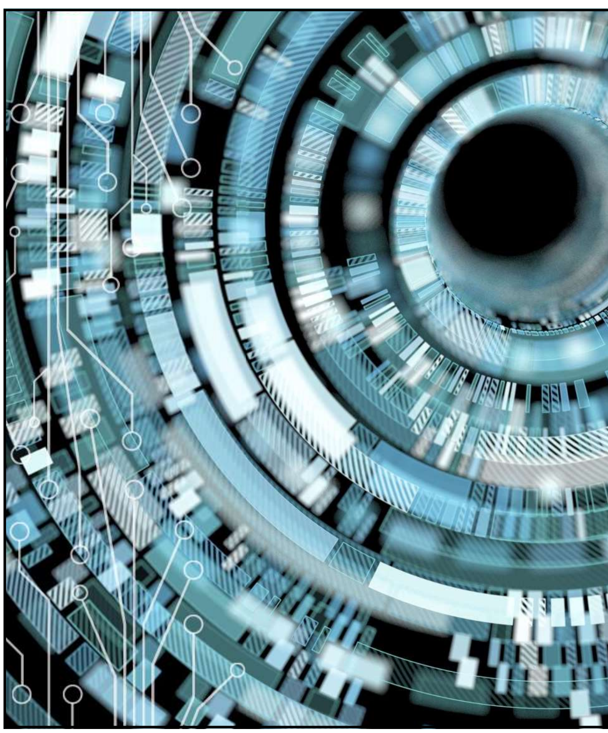


6



WHICH TYPE OF AI IS BEST FOR HEALTH CARE?

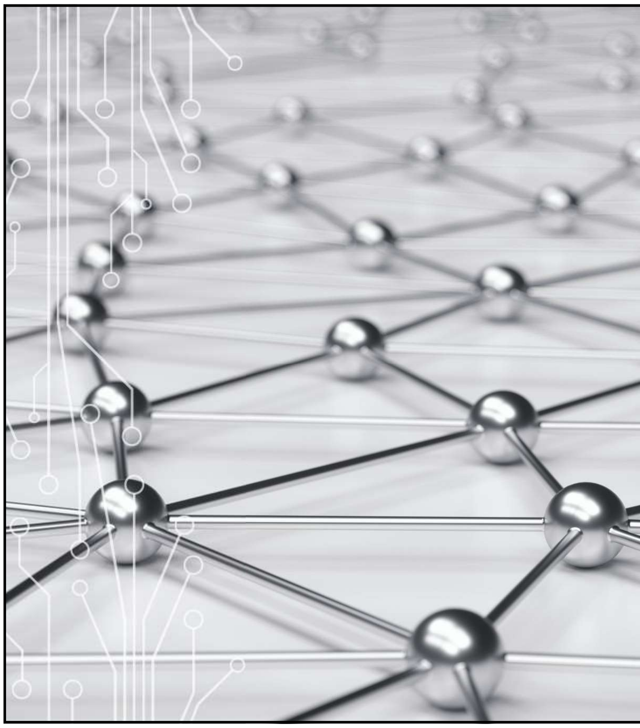
7



UNSUPERVISED MACHINE LEARNING

- Machine learns from huge, raw datasets
  - No preprocessing
  - No annotation or curation
  - No testing for bias or unintended outcomes
- Generative AI largely unsupervised
  - ChatGPT 3 (Generative Pretrained Transformer 3)
    - Trained on 570GB of data from text databased on Internet
  - Hallucinations, incorrect information

8



SUPERVISED MACHINE LEARNING

- Machines learn from curated, annotated datasets
- Humans do the curation and annotation
- Humans are involved in monitoring results and refining the models

9

IS ALL SUPERVISED MACHINE LEARNING EQUAL?

HOW MANY HAVE HAD A GOOD SUPERVISOR?

HOW MANY HAVE HAD A BAD SUPERVISOR?

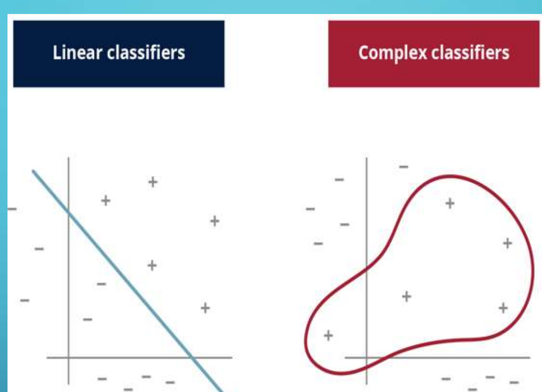
10

HOW DO WE BUILD SML?

- We identify "**features**" we think will delineate between two categories
  - Patients who are likely to develop cancer
    - Radiologic imaging, pathology report, diagnoses, family history, genetic data, etc.
- Humans label (**annotate**) those features
  - Cancer risk: grade of cancer, hormonal status of tumor, etc.
- Humans provide positive and negative cases (**training data set**)
  - Patients who were fine (negative) and patients who had poor outcomes (positive)
  - All features and labels (feature vectors) included with each case
- AI learns/is taught to classify cases based on patterns
  - What is the best (most predictive) and most parsimonious pattern?
  - Algorithm generates a "Decision Boundary" drawn around data features/vectors

11

IDENTIFYING BEST PATTERNS: DECISION BOUNDARIES



- Cases (+ and -) are plotted based on data vectors (collections of "features" on "layers")
- Algorithm creates boundaries to delineate positive and negative
- Boundaries can be simple (linear) or complex, based on feature vectors and layers

Figure from Barzilay lecture, MIT Sloan Course "AI in Health Care"

12

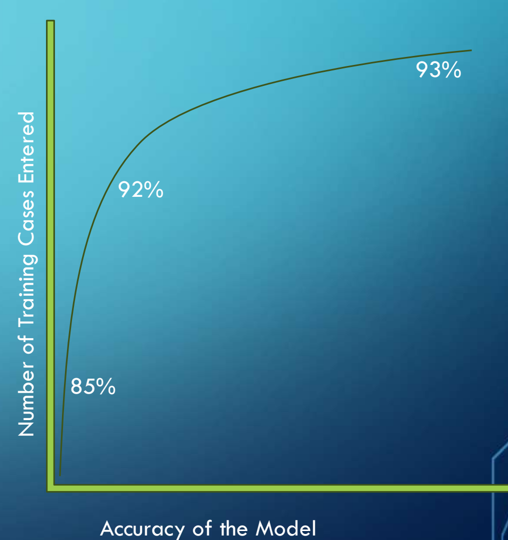
HOW DO WE BUILD SML?

- Area under the curve (AUC): how well the AI classifies negative and positive cases in the training data set
- Monitor and plot AUC as data are entered (learning curves)
- Provide NEW positive and negative case set (**validation set**)
  - Don't tell the AI which is which

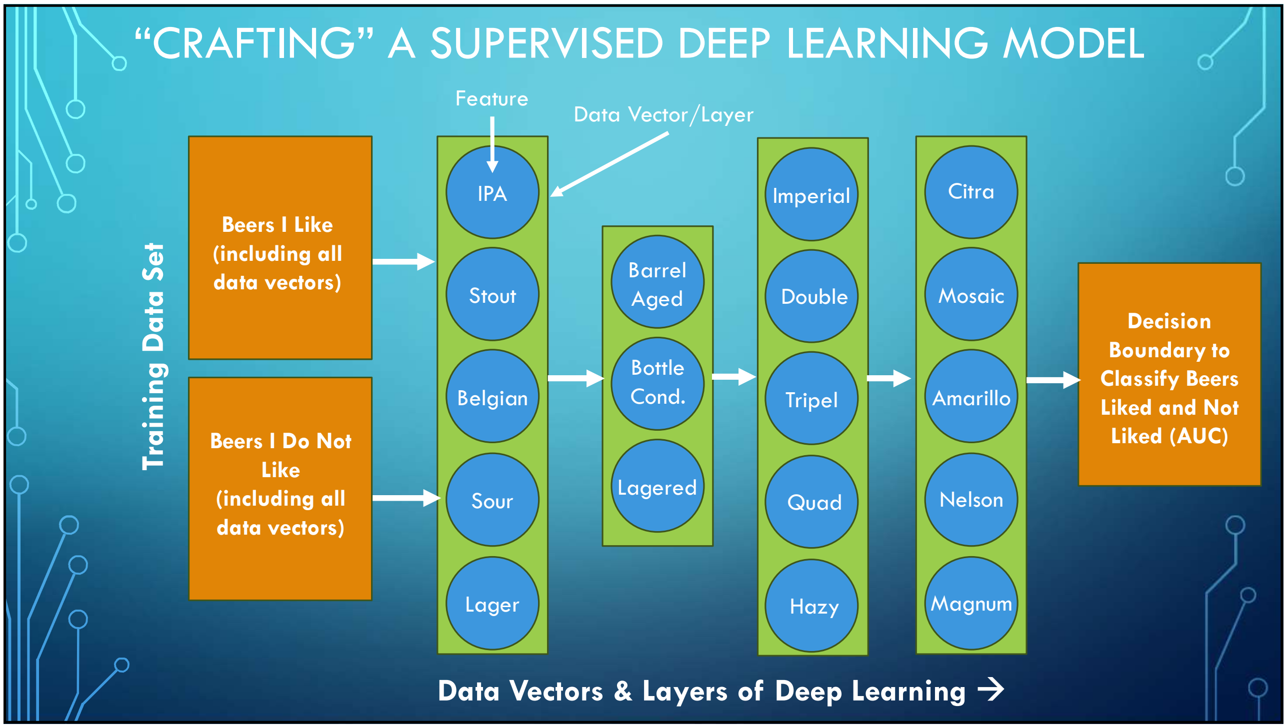
13

LEARNING CURVES AND DATA

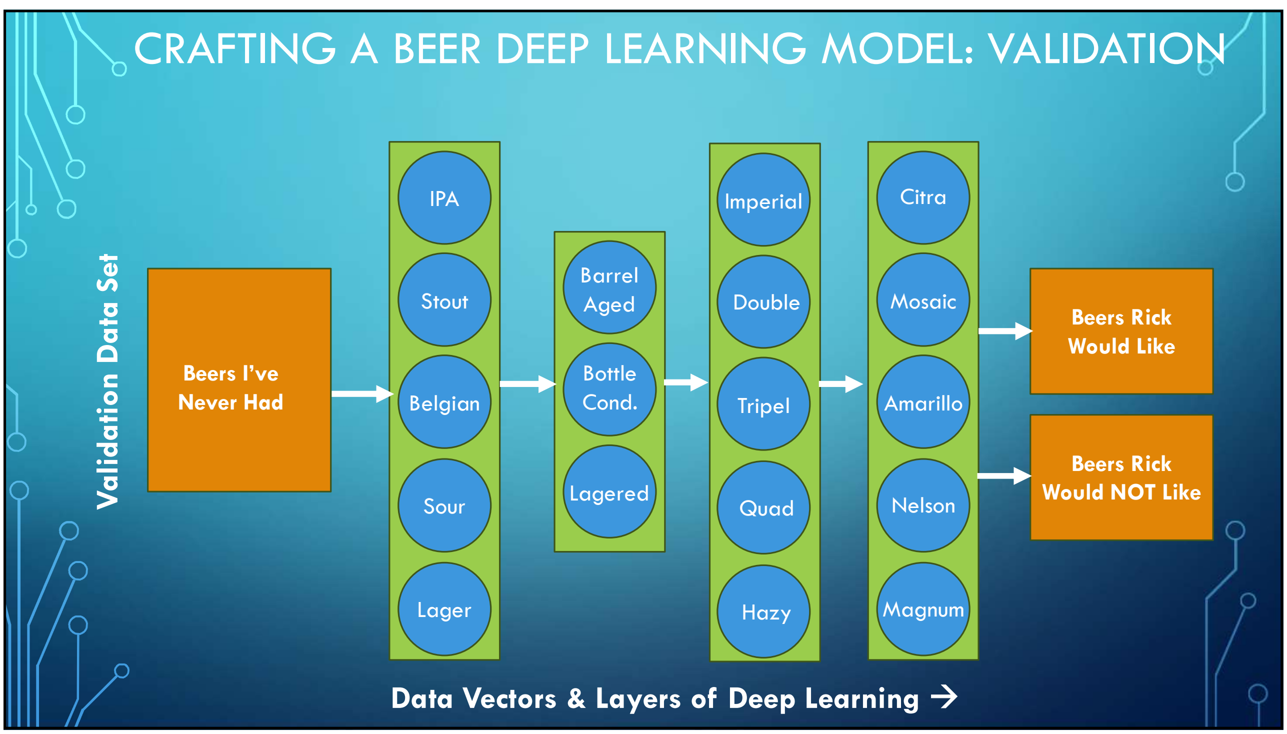
- Learning Curve
  - Number of cases (x axis) plotted against the AUC (y axis)
- AUC grows fastest from case 1–100 (~85%)
- Slower from 100–500 (~92%)
- Slower from 500–1000 (~93%)



14


"CRAFTING" A SUPERVISED DEEP LEARNING MODEL
Feature
Data Vector/Layer
Training Data Set
Beers I Like (including all data vectors)
Beers I Do Not Like (including all data vectors)
IPA
Stout
Belgian
Sour
Lager
Barrel Aged
Bottle Cond.
Lagered
Imperial
Double
Tripel
Quad
Hazy
Citra
Mosaic
Amarillo
Nelson
Magnum
Decision Boundary to Classify Beers Liked and Not Liked (AUC)
Data Vectors & Layers of Deep Learning →

15


CRAFTING A BEER DEEP LEARNING MODEL: VALIDATION
Validation Data Set
Beers I've Never Had
IPA
Stout
Belgian
Sour
Lager
Barrel Aged
Bottle Cond.
Lagered
Imperial
Double
Tripel
Quad
Hazy
Citra
Mosaic
Amarillo
Nelson
Magnum
Beers Rick Would Like
Beers Rick Would NOT Like
Data Vectors & Layers of Deep Learning →

16

Scientists turn to AI to make beer taste even better

Researchers in Belgium use artificial intelligence to improve taste, but say the skill of the brewer remains vital

The researchers constructed models based on machine learning – a form of AI – to predict how a beer would taste, and its appreciation, based on its composition. Photograph: Jack Andersen/Getty Images

https://www.theguardian.com/technology/2024/mar/26/ai-beer-taste-better-belgium-scientists

17

RISKS, MYTHS, AND LIMITATIONS OF AI

- Democratization
  - Cannot apply premade models with preconstructed algorithms to new situations.
  - Customization is key
  - Distributional shift: different geography, systems, and practices, limit generalizability
- AI and bias
  - AI does not introduce bias, it learns biases from biased data
  - If there are biases in your training data, your AI will learn those biases
  - Curation is critical
  - Train AI model to focus on subsets of the data to find different patterns
- Black-box concerns
  - Explainability is helping to address the concern
  - What happens when it is drastically better than humans but for reasons we cannot fully understand?

18

CUSTOMIZED AI SOLUTIONS

WHY CUSTOM, IN-HOUSE SUPERVISED MACHINE LEARNING IS THE FUTURE OF AI IN HEALTH CARE

19

REDUCING BURDEN, AUGMENTING CLINICAL WORKFLOWS, AND IMPROVING OUTCOMES

20

BENEFITS OF AI

- AI often seen as a barrier between doctor and patient
  - But can remove barriers and create time for more patient interaction
- Potential benefits
  - Remove burden of repetitive tasks
  - Augment patient workflows
  - Improve outcomes through customized care

21

EMERGENCY ROOM: CLINICAL GUIDELINES

- Beth Israel Deaconess Medical Center
  - David Sontag, Associate Professor of EE & Computer Science, Institute for Medical Engineering & Science
- When/with which patient should guidelines be applied?
- AI model to predict cardiac etiology
  - NLP to mine patient notes
  - Deep learning to identify most predictive factors
  - Connected to EHR
- Results
  - Instantly surfaces order sets relevant for patients with cardiac conditions
  - Recommends appropriate clinical guidelines
  - Human considers and applies

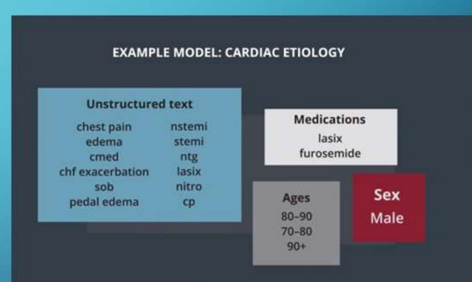


Image from David Sontag lecture, MIT course on AI in Health Care.

22

MEDICAL DOCUMENTATION AT POINT OF CARE

- Chief complaints
  - Can vary, which makes it harder to standardize
  - Many "downstream" tasks rely on chief complaint
    - Enrolling patients in clinical trials
    - Conducting retrospective QI studies
  - Take time to generate
- Increase standardization and save time?
- Redesigned workflow
  - Nurses normally assign chief complaint first
  - Instead, had nurse asks questions, write triage note
- AI processes note in real-time
  - Suggests chief complaint

23

EMERGENCY ROOM: MEDICAL DOCUMENTATION AT POINT OF CARE

Example

Writes triage note:	Enters vitals
• 78 y/o male c/o abdominal pain x3. c/o nausea and vomiting.	• BP = 120/75, O2sat = 99%, temp = 99.1, HR = 105, Respirator Rate = 16

↓

Option 1: click on one of five suggested chief complaints

↓

Option 2: begin typing in free text field and selected from list of auto-completes, presented in order of probability

24

ER: MEDICAL DOCUMENTATION AT POINT OF CARE

- Option 2: Free text entry with contextual autocomplete
  - Sorted by most probable chief complaint, according to the AI

Image from David Sontag lecture, MIT course on AI in Health Care.

25

ER: CHIEF COMPLAINT STANDARDIZATION

- 60,000 patients per year
- Initially, only 20-30% of free text chief complaints could be standardized
- After 4 years, was nearly 100%

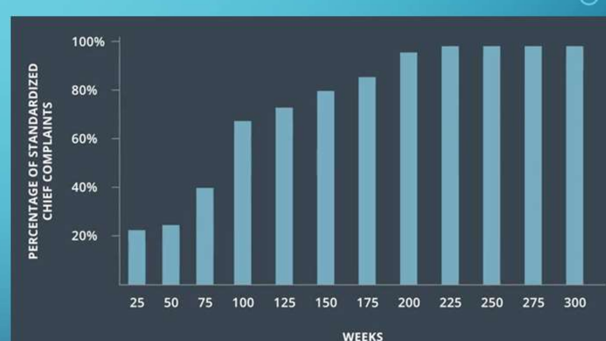


Image from David Sontag lecture, MIT course on AI in Health Care.

26

ER: CHIEF COMPLAINT TIME SAVINGS

- Initially took 11.6 keystrokes to enter chief complaint
- After AI use, took 0.6 keystrokes
  - Sometimes chief complaint was in the top five listed, so no typing in the text field at all

27

AI VS "USUAL CARE" IN ER SETTINGS

- Systematic review
- 23 of 1,656 studies selected, representing 16,274,647 patients
  - Diagnostic studies (n=7) showed AI outperformed usual care in all performance metrics
  - In-hospital mortality studies (n=6) best-performing AI had better discrimination (.74-.94) than any other clinical decision tool (.68-.81)
  - Hospitalization studies (n=4) showed AI had better discrimination (.80-.83) than triage-based scores (.69-.82)

Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review. Acad Emerg Med. 2021;28(2):184-196. doi:10.1111/acem.14190

28

IMPROVING PATIENT CARE WITH AI AND IMAGING

29

BREAST CANCER AND AI

- Dr. Connie Lehman
  - Director of Breast Imaging and Co-Director of the Avon Comprehensive Breast Evaluation Center, Massachusetts General
- Breast density
  - Federal mandate to inform women with high density
  - Higher risk of tumors, tumors are harder to detect
- Radiologists are inconsistent
  - Classifications of high density ranged from 6% to 85% for same set of images

30

BREAST CANCER AND AI

- Built AI to read mammograms and assign density rating
- Integrated into clinical workflow at Massachusetts General
  - Better and more consistent performance than humans
- Radiologist had the final say to accept or reject AI reading
  - Learned intermediary
- All rejections by radiologist sent to expert panel of radiologists
- Panel determinations were nearly always in agreement with AI

31

RISK STRATIFICATION: BREAST CANCER

- Images contain billions of data points
  - Pixels per image
  - Multiple images from different angles
- Most of that data is not used
  - Condensed into 1-2 page summary
  - Further condensed into a few categories
    - Cancer grade, hormonal status
- 4 women with same categorization
- 3 were fine, 1 metastatic recurrence

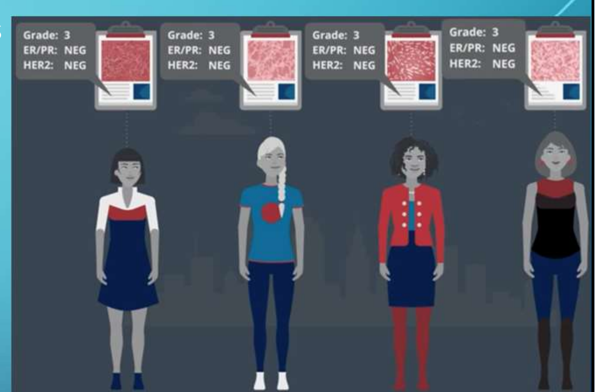


Image from lecture by Regina Barzilay, MIT Delta Electronics Professor of EE and Computer Science, course on AI in Health Care.

32

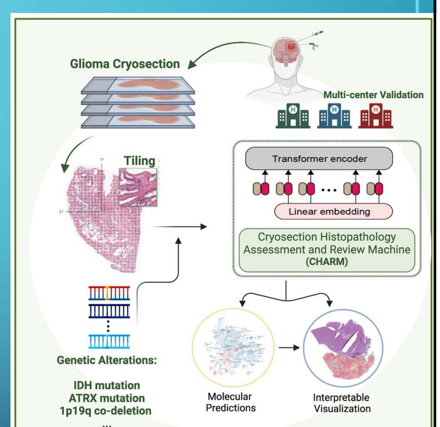
RISK STRATIFICATION: BREAST CANCER

- Standard risk assessment is Tyrer-Cuzik
  - AUC is .64 overall, .66 for one year, .65 for two year, and .63 for five year
  - UNLESS you are Black or Asian, then it is .58 AND .53 at five years
- Deep learning AI
  - Considered all images and patient data/outcomes over time
    - Data vector for each point in time
    - Features for things like change since last result
  - Annotated cases with treatments and outcomes over time
- AI better than humans
  - AUC is .88 for one year, .79 for two year, and .79 for five year
  - .72 for Black women and .76 for Asian women

33

INTRAOPERATIVE TUMOR CRYOSECTION EVALUATION

- Problem
  - Surgery for brain cancer requires tumor classification
  - Tissue freezing makes identification hard (artifacts)
  - Tumor classification includes molecular profiles
- Solution
  - Cryosection Histopathology Assessment & Review Machine
  - 1,524 glioma patients
  - Validation at three different centers
- Results
  - Identified malignant cells (AUC .98), IDH-mutant vs. wild (AUC .79-.82), three types of molecularly defined gliomas (AUC .88-.93), and most prevalent IDH-mutant tumors (AUC .89-.97).



Nasrallah MP, Zhao J, Tsai CC, et al. Machine learning for cryosection pathology predicts the 2021 WHO classification of glioma. Med. 2023;4(8):526-540.e4. doi:10.1016/j.medj.2023.06.002

34

AGE-RELATED MACULAR DEGNERATION

- Problem one
  - AMD progression is atypical (less than 5 years) in 10-20% of patients
  - Risk stratification/prediction is critical and requires significant clinical expertise
  - Shortage of expertise and high variation on classifications
- Solution one: iPredict
  - Deep learning model lets non-eye care specialists screen for AMD and predict risk
  - Trained on 93,380 color fundus photos from 4,757 participants in AREDS 10-year study
  - 1,824 images from individuals who had accelerated progression; 2,840 images from those who did not
  - Validated on 23,495 images

35

AGE-RELATED MACULAR DEGNERATION

- Solution one: iPredict
  - In less than 60 seconds, model classified patient as referable or not for AMD (AUC .99)
  - Predicts risk of developing late AMD within 2 years (AUC .84–.86)
    - Bhuiyan A, Wong TY, Ting DSW, Govindaiah A, Souied EH, Smith RT. Artificial Intelligence to Stratify Severity of Age-Related Macular Degeneration (AMD) and Predict Risk of Progression to Late AMD. Transl Vis Sci Technol. 2020;9(2):25. Published 2020 Apr 24. doi:10.1167/tvst.9.2.25
- Solution two: NIH system
  - 3,298 participants, 80,000 images
  - Outperformed the accuracy of retinal specialists using two clinical standards
    - Peng Y, Keenan TD, Chen Q, et al. Predicting risk of late age-related macular degeneration using deep learning. NPJ Digit Med. 2020;3:111. Published 2020 Aug 27. doi:10.1038/s41746-020-00317-z

36

AI AND PATIENT SAFETY OUTCOMES

- Systematic Literature Review of AI and Patient Safety Outcomes
  - Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform. 2020;8(7):e18599. Published 2020 Jul 24. doi:10.2196/18599
- 53 studies showed improvements
  - Most from 85% to 95% AUC
  - All with practical significance

37

Chen L, Dubrawski A, Wang D, Fiterau M, Guillame-Bert M, Bose E, et al. Using Supervised Machine Learning to Classify Real Alerts and Artifact in Online Multisignal Vital Sign Monitoring Data. Crit Care Med 2016 Jul;44(7):e456-e463 doi: 10.1097/CCM.0000000000001660]	Machine-learning (ML) models could distinguish clinically relevant pulse arterial O2 saturation, blood pressure, and respiratory rate from artifacts in an online monitoring dataset (AUC>0.87)
Ansari S, Belle A, Ghanbari H, Salamango M, Najarian K. Suppression of false arrhythmia alarms in the ICU: a machine learning approach. Physiol Meas 2016 Aug;37(8):1186-1203. doi: 10.1088/0967-3334/37/8/1186	ML algorithm along with MMD was effective in suppressing false alarms
Zhang Q, Chen X, Fang Z, Zhan Q, Yang T, Xia S. Reducing false arrhythmia alarm rates using robust heart rate estimation and cost-sensitive support vector machines. Physiol Meas 2017 Feb;38(2):259-271. doi: 10.1088/1361-6579/38/2/259	SVM reduced false alarm rates. The model gave an overall true positive rate of 95% and true negative rate of 85%
Antink CH, Leonhardt S, Walter M. Reducing false alarms in the ICU by quantifying self-similarity of multimodal biosignals. Physiol Meas 2016 Aug;37(8):1233-1252. doi: 10.1088/0967-3334/37/8/1233	A false alarm reduction score of 65.52 was achieved; employing an alarm-specific strategy, the model performed at a true positive rate of 95% and true negative rate of 78%. False alarms for extreme tachycardia were suppressed with 100% sensitivity and specificity
Eerikäinen LM, Vanschoren J, Rooijakkers MJ, Vullings R, Aarts RM. Reduction of false arrhythmia alarms using signal selection and machine learning. Physiol Meas 2016 Aug 25;37(8):1204-1216. doi: 10.1088/0967-3334/37/8/1204	Out of 5 false alarms, 4 were suppressed; 77.39% real-time model accuracy
Ménard T, Barmaz Y, Koneswarakantha B, Bowling R, Popko L. Enabling Data-Driven Clinical Quality Assurance: Predicting Adverse Event Reporting in Clinical Trials Using Machine Learning. Drug Saf 2019 Sep 23;42(9):1045-1053 doi: 10.1007/s40264-019-00831-4	The ML method identified the sites by risk of underreporting and enabled real-time safety reporting. The proposed model had an AUC of 0.62, 0.79, and 0.92 for simulation scenarios of 25%, 50%, and 75%, respectively. This project was part of a broader effort at Roche/Genentech to augment and complement traditional clinical quality assurance approaches
Segal G, Segev A, Brom A, Lifshitz Y, Wasserstrum Y, Zimlichman E. Reducing drug prescription errors and adverse drug events by application of a probabilistic, machine-learning based clinical decision support system in an inpatient setting. J Am Med Inform Assoc 2019 Dec 01;26(12):1560-1565. doi: 10.1093/jamia/ocz135	85% of the alerts were clinically valid, and 80% were considered clinically useful; 43% of the alerts caused changes in subsequent medical orders. Thus, the model detected medication errors
Hu SB, Wong DJL, Correa A, Li N, Deng JC. Prediction of Clinical Deterioration in Hospitalized Adult Patients with Hematologic Malignancies Using a Neural Network Model. PLoS One 2016;11(8):e0161401 doi: 10.1371/journal.pone.0161401	NN-based model could detect health deterioration such as heart rate variability with more accuracy than one of the best-performing early warning scores (ViEWS). The positive prediction value of NN was 77.58% and the negative prediction value was 99.19%
Kwon J, Lee Y, Lee Y, Lee S, Park J. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. J Am Heart Assoc 2018 Jun 26;7(13):e008678 doi: 10.1161/JAHA.118.008678	The DEWS identified more than 50% of patients with in- hospital cardiac arrest 14 hours before the event. It allowed medical staff to have enough time to intervene. The AUC and AUPRC of DEWS was 0.85 and 0.04, respectively, and outperformed MEWS with AUC and AUPC of 0.60 and 0.003, respectively; RF with AUC and AUPC of 0.78 and 0.01, respectively; and LR with AUC and AUPRC of 0.61 and 0.007, respectively. DEWS reduced the number of alarms by 82.2%, 13.5%, and 42.1% compared with the other models at the same sensitivity
Gupta J, Patrick J. Automated validation of patient safety clinical incident classification: macro analysis. Stud Health Technol Inform 2013;188:52-57.	The selected models performed poorly in classifying incident categories (48.77% best, using J48), but performed comparatively better in classifying free text (76.49% using NB).

38

Wang Y, Coiera EW, Runciman W, Magrabi F. Automating the Identification of Patient Safety Incident Reports Using Multi-Label Classification. : IOS Press; 2017 Presented at: Precision Healthcare Through Informatics: Proceedings of the 16th World Congress on Medical and Health Informatics; August 21-25, 2017; Hangzhou, China p. 609-613.	Binary classifier improved identification of common incident types: falls, medications, pressure injury, aggression, documentation problem, and others. Automated identification enabled safety problems to be detected and addressed in a more timely manner
Fong A, Harriott N, Walters DM, Foley H, Morrissey R, Ratwani RR. Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events. Int J Med Inform 2017 Aug;104:120-125. doi: 10.1016/j.ijmedinf.2017.05.005	ML algorithms identified the medication event originating stages, event types, and causes, respectively. The models improved the efficiency of analyzing the medication event reports and learning from the reports in a timely manner with (SVM) F1 of 0.792 and (RF) F1 of 0.925
ElMessiry A, Zhang Z, Cooper W, Catron T, Karrass J, Singh M, editors. Leveraging sentiment analysis for classifying patient complaints. 2017 Presented at: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,Health Informatics; 2017; Boston. doi: 10.1145/3107411.3107421	Care-related complaints were influenced by money and emotion
Chondrogiannis E, Andronikou V, Varvarigou T, Karanastasis E, editors. Semantically-Enabled Context-Aware Abbreviations Expansion in the Clinical Domain. 2017 Presented at: Proceedings of the 9th International Conference on Bioinformatics Biomedical Technology; 2017; Washington DC. doi: 10.1145/3093293.3093304	Each clinical study document contained about 6.8 abbreviations. Each abbreviation can have 1.25 meanings on average. This helped in identification of acronyms
Liang C, Gong Y. Automated Classification of Multi-Labeled Patient Safety Reports: A Shift from Quantity to Quality Measure. Stud Health Technol Inform 2017;245:1070-1074.	Binary relevance was the best problem transformation algorithm in the multilabeled classifiers. It provided suggestions on how to implement automated classification of patient safety reports in clinical settings
Ong M, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc 2012 Jun;19(e1):e110-e118 doi: 10.1136/amiajnl-2011-000562	SVM performed well on datasets with diverse incident types (85.8%) and data with patient misidentification (96.4%). About 90% of false positives were found in "near-misses" and 70% of false negative occurred due to spelling errors
Taggart M, Chapman WW, Steinberg BA, Ruckel S, Pregenzer-Wenzler A, Du Y, et al. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. JAMA Netw Open 2018 Oct 05;1(6):e183451 doi: 10.1001/jamanetworkopen.2018.3451	Rule-based NLP was better than the ML approach. NLP detected bleeding complications with 84.6% specificity, 62.7% positive predictive value, and 97.1% negative predictive value. It can thus be used for quality improvement and prevention programs
Denecke K, Lutz HS, Pöpel A, May R, editors. Talking to ana: A mobile self-anamnesis application with conversational user interface. 2018 Presented at: Proceedings of the 2018 International Conference on Digital Health; 2018; Lyon. doi: 10.1145/3194658.3194670	Electronic health platform provides an intuitive conversational user interface that patients use to connect to their therapist and self-anamnesis app. The app also allows data sharing among treating therapists
Evans HP, Anastasiou A, Edwards A, Hibbert P, Makeham M, Luz S, et al. Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. Health Informatics J 2019 Mar 07:1460458219833102. doi: 10.1177/1460458219833102	The SVM classifier improved the identification of patient safety incidents. Incident reports containing deaths were most easily classified with an accuracy of 72.82%. The severity classifier was not accurate to replace manual scrutiny
Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. J Am Med Inform Assoc 2019 Dec 01;26(12):1600-1608. doi: 10.1093/jamia/ocz146	CNN achieved high F scores (>85%) across all test datasets when identifying common incident types, including falls, medications, pressure injury, and aggression. It improved the process by 11.9% to 45.10% across different datasets

39

Li M, Ladner D, Miller S, Classen D. Identifying hospital patient safety problems in real-time with electronic medical record data using an ensemble machine learning model. Int J Clin Med Inform 2018;1(1):43-58.	The adverse event risk score at the 0.1 level could identify 57.2% of adverse events with 26.3% accuracy from 9.2% of the validation sample. The adverse event risk score of 0.04 could identify 85.5% of adverse events
Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. doi: 10.1001/jama.2011.1204	NLP identified 82% of acute renal failure cases compared with 38% for patient safety indicators. Similar results were obtained for venous thromboembolism (59% vs 46%), pneumonia (64% vs 5%), sepsis (89% vs 34%), and postoperative myocardial infarction (91% vs 89%)
Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak 2017 Jun 12;17(1):84 doi: 10.1186/s12911-017-0483-8	For severity level, the F score for severity assessment code (SAC) 1 (extreme risk) was 87.3 and 64% for SAC4 (low risk) on balanced data. With stratified data, a high recall was achieved for SAC1 (82.8%-84%), but precision was poor (6.8%-11.2%). High-risk incidents (SAC2) and medium-risk incidents (SAC3) were often misclassified. Reports about falls, medications, pressure injury, aggression, and blood tests were identified with high recall and precision
Rosenbaum M, Baron J. Using Machine Learning-Based Multianalyte Delta Checks to Detect Wrong Blood in Tube Errors. Am J Clin Pathol 2018 Oct 24;150(6):555-566. doi: 10.1093/ajcp/aqy085	In contrast to the univariate analysis, the best performing multivariate delta check model (SVM) identified errors with a high degree of accuracy (0.97)
McKnight SD. Semi-supervised classification of patient safety event reports. J Patient Saf 2012 Jun;8(2):60-64. doi: 10.1097/PTS.0b013e31824ab987	The semisupervised model categorized patient safety reports into their appropriate patient safety topic and avoided over- laps; 85% of unlabeled reports were assigned correct labels. It helped NCPS analysts to develop policy and mitigation decisions
Marella WM, Sparnon E, Finley E. Screening Electronic Health Record-Related Patient Safety Reports Using Machine Learning. J Patient Saf 2017 Mar;13(1):31-36. doi: 10.1097/PTS.0000000000000104	The NB kernel performed best, with an AUC of 0.927, accu- racy of 0.855, and F score of 0.877. The overall proportion of cases found relevant was compara- ble between manually and automatically screened cases; 334 reports identified by the model as relevant were identified as not relevant, implying a false-positive rate of 13%. Manual screening identified 4 incorrect predictions, implying a false-negative rate of 29%
Ye C, Wang O, Liu M, Zheng L, Xia M, Hao S, et al. A Real-Time Early Warning System for Monitoring Inpatient Mortality Risk: Prospective Study Using Electronic Medical Record Data. J Med Internet Res 2019 Jul 05;21(7):e13719 doi: 10.2196/13719	The modified early warning system accurately predicted the possibility of death for the top 13.3% (34/255) of patients at least 40.8 hours before death
Fong A, Adams KT, Gaunt MJ, Howe JL, Kellogg KM, Ratwani RM. Identifying health information technology related safety event reports from patient safety event report databases. J Biomed Inform 2018 Oct;86:135-142 doi: 10.1016/j.jbi.2018.09.007	Unigram models performed better than Bigram and combined models. It identified HIT-related events trained on PSE free-text descriptions from multiple states and health care systems. The unigram LR model gave an AUC of 0.931 and an F1 score of 0.765. LR also showed potential to maintain a faster runtime when more reports are analyzed. The final HIT model had less complexity and was more easily sharable
Simon ACR, Holleman F, Gude WT, Hoekstra JBL, Peute LW, Jaspers MWM, et al. Safety and usability evaluation of a web-based insulin self-titration system for patients with type 2 diabetes mellitus. Artif Intell Med 2013 Sep;59(1):23-31. doi: 10.1016/j.artmed.2013.04.009	27 out of 74 (36.5%) PANDIT advice differed from those provided by diabetes nurses. However, only one of these (1.4%) was considered unsafe by the panel
Song D, Chen Y, Min Q, Sun Q, Ye K, Zhou C, et al. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. J Clin Pharm Ther 2019 Apr 18;44(2):268-275. doi: 10.1111/jcpt.12786	The 10-fold crossvalidation improved the identification of drug-drug interaction with AUC>0.97, which is significantly greater than the analogously developed ML model (0.67)

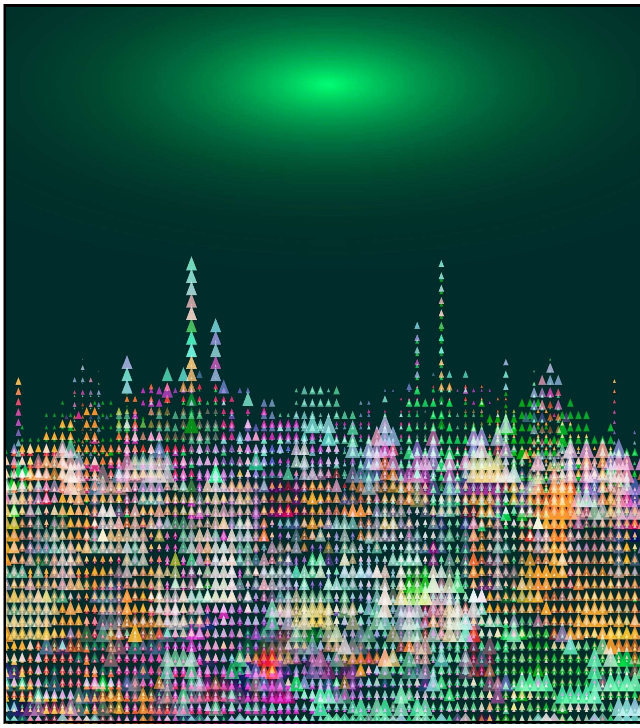
40

Hammann F, Gutmann H, Vogt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. Clin Pharmacol Ther 2010 Jul 10;88(1):52-59. doi: 10.1038/clpt.2009.248	CART exhibited high predictive accuracy of 78.94% for allergic reactions, 88.69% for renal, and 90.22% for the liver. CHAID model showed a high accuracy of 89.74% for the central nervous system
Bean DM, Wu H, Iqbal E, Dzahini O, Ibrahim ZM, Broadbent M, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Sci Rep 2017 Nov 27;7(1):16416. doi: 10.1038/s41598-017-16674-x	The proposed model (own model) outperformed traditional LR, SVM, DT, and predicted adverse drug reactions with an AUC of 0.92
Hu Y, Tai C, Tsai C, Huang M. Improvement of Adequate Digoxin Dosage: An Application of Machine Learning Approach. J Healthc Eng 2018;2018:3948245. doi: 10.1155/2018/3948245	In the non drug-drug interaction group, the AUC of RF, MLP, CART, and C4.5 was 0.91, 0.81, 0.79, and 0.784, respectively; for the drug-drug interaction group, the AUC of RF, CART, MLP, and C4.5 was 0.89, 0.79, 0.77, and 0.77, respectively. DT-based approaches and MLP can determine the initial dosage of a high-alert digoxin medication, which can increase drug safety in clinical practice
Tang Y, Yang J, Ang PS, Dorajoo SR, Foo B, Soh S, et al. Detecting adverse drug reactions in discharge summaries of electronic medical records using Readpeer. Int J Med Inform 2019 Aug;128:62-70. doi: 10.1016/j.ijmedinf.2019.04.017	A total of 33 trial sets were evaluated by the algorithm and reviewed by pharmacovigilance experts. After every 6 trial sets, drug and adverse event dictionaries were updated, and rules were modified to improve the system. The model identified adverse events with 92% precision and recall
Hu Y, Wu F, Lo C, Tai C. Predicting warfarin dosage from clinical data: a supervised learning approach. Artif Intell Med 2012 Sep;56(1):27-34. doi: 10.1016/j.artmed.2012.04.001	The proposed model improved warfarin dosage when com- pared to the baseline (mean absolute error 0.394); reduced mean absolute error by 40.04%
Hasan S, Duncan GT, Neill DB, Padman R. Automatic detection of omissions in medication lists. J Am Med Inform Assoc 2011 Jul 01;18(4):449-458 doi: 10.1136/amiajnl-2011-000106	Collaborative filtering identified the top 10 missing drugs about 40% to 50% of the time and the therapeutic missing drugs about 50% to 65% of the time
Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using Artificial Intelligence to Reduce the Risk of Nonadherence in Patients on Anticoagulation Therapy. Stroke 2017 May;48(5):1416-1419 doi: 10.1161/STROKEAHA.116.016281	Mean (SD) cumulative adherence based on the AI platform was 90.5% (7.5%). Plasma drug concentration levels indicated that adherence was 100% (15/15) and 50% (6/12) in the intervention and control groups, respectively
Long J, Yuan MJ, Poonawala R. An Observational Study to Evaluate the Usability and Intent to Adopt an Artificial Intelligence-Powered Medication Reconciliation Tool. Interact J Med Res 2016 May 16;5(2):e14 doi: 10.2196/ijmr.5462	All patients completed the task. The software improved reconciliation; all patients identified at least one error in their electronic medical record medication list; 8 of 10 patients reported that they would use the device in the future. The entire team (clinical and patients) liked the device and preferred to use it in the future
Reddy M, Pesl P, Xenou M, Toumazou C, Johnston D, Georgiou P, et al. Clinical Safety and Feasibility of the Advanced Bolus Calculator for Type 1 Diabetes Based on Case-Based Reasoning: A 6-Week Nonrandomized Single-Arm Pilot Study. Diabetes Technol Ther 2016 Aug;18(8):487-493. doi: 10.1089/dia.2015.0413	ABC4D was safe for use as an insulin bolus dosing system. A trend suggesting a reduction in postprandial hypoglycemia was observed. The median (IQR) number of postprandial hypoglycemia episodes within 6 h after the meal was 4.5 (2.0-8.2) in week 1 versus 2.0 (0.5-6.5) in week 6 (P=.10). No episodes of severe hypoglycemia occurred during the study
Schiff GD, Volk LA, Volodarskaya M, Williams DH, Walsh L, Myers SG, et al. Screening for medication errors using an outlier detection system. J Am Med Inform Assoc 2017 Mar 01;24(2):281-287. doi: 10.1093/jamia/ocw171	75% of the chart-reviewed alerts generated by MedAware were valid from which medication errors were identified. Of these valid alerts, 75.0% were clinically useful in flagging potential medication errors.

41

Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 06;15:37 doi: 10.1186/s12911-015-0160-8	The hybrid algorithm yielded precision (P) of 95.0%, recall (R) of 91.6%, and F value of 93.3% on medication entity identification, and P=98.7%, R=99.4%, and F=99.1% on attribute linkage. The combination of the hybrid system and medication matching system gave P=92.4%, R=90.7%, and F=91.5%, and P=71.5%, R=65.2%, and F=68.2% on classifying the matched and the discrepant medications, respectively
Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform 2015 Dec;84(12):1057-1064. doi: 10.1016/j.ijmedinf.2015.09.002	The NLP-assisted manual review identified an additional 728 (3.1%) patients with evidence of clinically diagnosed problem opioid use in clinical notes.
Tinoco A, Evans RS, Staes CJ, Lloyd JF, Rothschild JM, Haug PJ. Comparison of computerized surveillance and manual chart review for adverse events. J Am Med Inform Assoc 2011;18(4):491-497 doi: 10.1136/amiajnl-2011-000187	CSS detected more hospital-associated infections than man- ual chart review (92% vs 34%); CSS missed events that were not stored in a coded format
Onay A, Onay M, Abul O. Classification of nervous system withdrawn and approved drugs with ToxPrint features via machine learning strategies. Comput Methods Programs Biomed 2017 Apr;142:9-19. doi: 10.1016/j.cmpb.2017.02.004	The Gaussian SVM model yielded 78% prediction accuracy for the drug dataset, including all diseases. The ensemble of bagged tree and linear SVM models in- volved 89% of the accuracies for psycholeptics and psycho- analytic drugs
Cai R, Liu M, Hu Y, Melton BL, Matheny ME, Xu H, et al. Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. Artif Intell Med 2017 Feb;76:7-15 doi: 10.1016/j.artmed.2017.01.004	CARD demonstrated higher accuracy in identifying known drug interactions compared to the traditional method (20% vs 10%); CARD yielded a lower number of drug combinations that are unknown to interact (50% for CARD vs 79% for association rule mining).
Dandala B, Joopudi V, Devarakonda M. Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks. Drug Saf 2019 Jan;42(1):135-146. doi: 10.1007/s40264-018-0764-x	Joint modeling improved the identification of adverse drug events from 0.62 to 0.65
Dey S, Luo H, Fokoue A, Hu J, Zhang P. Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics 2018 Dec 28;19(Suppl 21):476 doi: 10.1186/s12859-018-2544-0	Neural fingerprints from the deep learning model (AUC=0.72) outperformed all other methods in predicting adverse drug reactions. The model identified important molecular substructures that are associated with specific adverse drug reactions
Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes. Drug Saf 2019 Jan;42(1):123-133 doi: 10.1007/s40264-018-0761-0	MADEx achieved the top-three best performances (F1 score of 0.8233) for clinical name entity recognition, adverse drug effect, and relations from clinical texts, which outperformed traditional methods
Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting Adverse Drug Events with Rapidly Trained Classification Models. Drug Saf 2019 Jan 16;42(1):147-156 doi: 10.1007/s40264-018-0763-y	The micro-averaged F1 score was 80.9% for named entity recognition, 88.1% for relation extraction, and 61.2% for the integrated systems
Lian D, Khoshneshin M, Street WN, Liu M. Adverse drug effect detection. IEEE J Biomed Health Inform 2013 Mar;17(2):305-311. doi: 10.1109/TITB.2012.2227272	Experimental results showed the usefulness of the proposed pattern discovery method by improving the standard baseline adverse drug reaction by 23.83%
Huang L, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC Genomics 2011 Dec 23;12 Suppl 5:S11 doi: 10.1186/1471-2164-12-S5-S11	The proposed computational framework showed that an in silico model built on this framework can achieve satisfactory cardiotoxicity adverse drug reaction prediction performance (median AUC=0.771, accuracy=0.675, sensitivity=0.632, and specificity=0.789).

42



HOSPITAL OPTIMIZATION TO IMPROVE CAPACITY AND OUTCOMES

43

HOSPITAL OPTIMIZATION

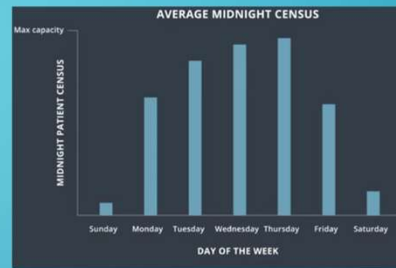
- Dimitris Bertsimas, Professor of Operations Research, MIT
  - Operating room optimization
    - Half of beds in hospital come from ER admissions, half from elective surgeries
    - Optimize operating rooms to maximize capacity for ER patients
  - Predict length of stay in hospital (discharge prediction)
  - Predict number of patients admitted via ER
- Beth Israel Deaconess Medical Center
- Use deep learning to look at all three and optimize patient flow

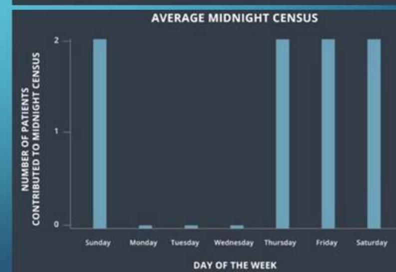
Bertsimas D and Pauphilet J. Forthcoming. "Holistic Hospital Optimization." *Management Science*.

44

OR SCHEDULING

- Peak bed usage on Wed./Thur.
- Surgeon X
  - 2 complex cases on Monday
    - 4-day length of stay
    - Contributes 2 patients to M–R census
  - N Ambulatory cases on Thursday
    - Same day discharge
- Switch schedule
  - Complex cases on Thursday
    - Stays R–M
  - Ambulatory on Monday (no stays)
    - Contributes no patients M–W





Images from Bertsimas lecture, MIT course on AI in Health Care. Data from Hartford Hospital.

45

OR SCHEDULING

- Surgeons tend to operate in patterns/blocks
  - Optimizing by assigning surgeons to different days
- Constraints used for their AI
  - Not all surgeons and surgeries can move
  - Number of surgeries cannot exceed previous year
  - Limited number of changes to schedule
  - No OR blocks on weekends
  - OR limited to 7 hours per day

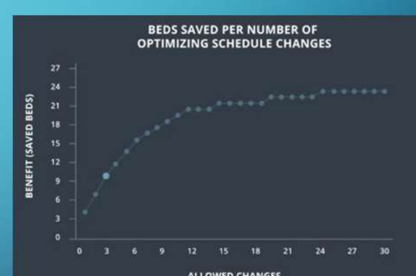
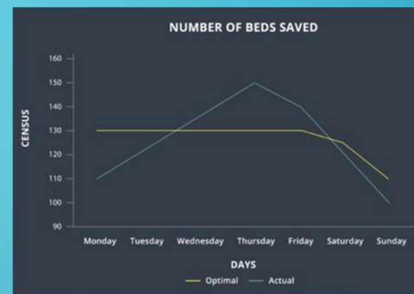


Image from Bertsimas lecture, MIT course on AI in Health Care. Data from Hartford Hospital.

46

OR SCHEDULING

- Implementation
  - AI generates multiple solutions in seconds
  - Solutions are approved/selected
- Results
  - Changed 11 out of 250 surgeons
  - Freed up 21 beds per week
- Most benefit comes from the first few iterations



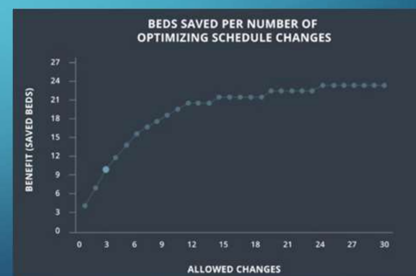


Image from Bertsimas lecture, MIT course on AI in Health Care. Data from Hartford Hospital.

47

PREDICTING SHORT-TERM DISCHARGE

- Beth Israel Deaconess Medical Center
- Used patient admits over 2.5 years
  - Excluded psych, OBGYN, newborns
  - 60,000 admissions, 40,000 patients
- NLP and Deep Learning
  - NLP: Used patient notes day after admittance (curated, annotated)
  - Deep Learning: Multiple feature vectors, including movement, walking, restricted to bed dietary, fall risk, comorbidities, vital signs, labs, medication, insurance, secure homesite, language, etc.
- Asked four questions
  - Discharged in 24 or 48 hours?
  - ICU in next 24 hours?
  - Exceed 7 or 14 day stay?
  - Was discharge location home, hospice, rehab, or death?

48

PREDICTING SHORT-TERM DISCHARGE

- AUC for discharge within 24 hours was between 81% and 84%
- ICU prediction was 97%
- Used to schedule staffing and resources

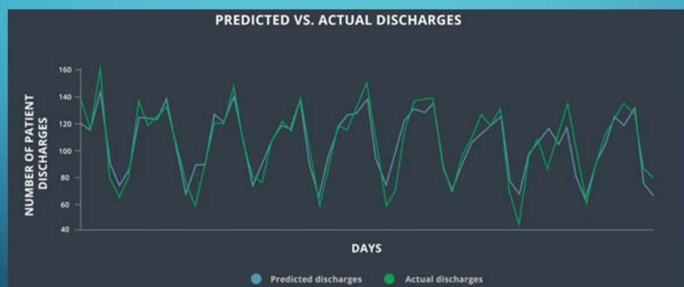


Image from Bertsimas lecture, MIT course on AI in Health Care. Data from Hartford Hospital.

49

THE INVISIBLES

50

51

EMERALD DEVICE

- Like Wi-Fi router
- Measures disturbances in electromagnetic waves throughout the home
- AI models interpret the patterns
- Breathing and heart rate
- As accurate as in-office measurement
- With more validity (in-situ)

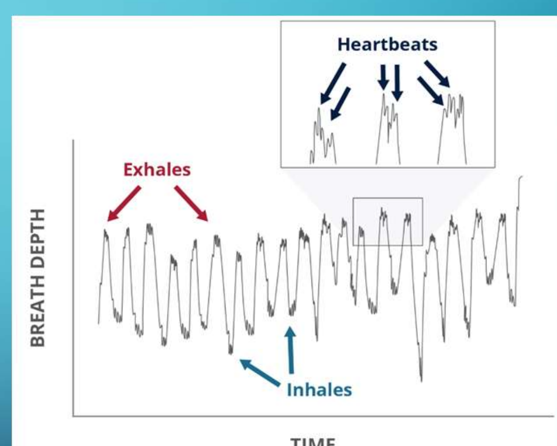
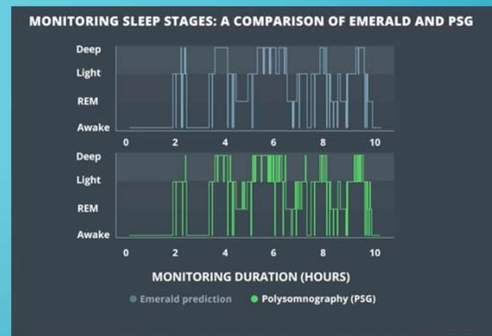


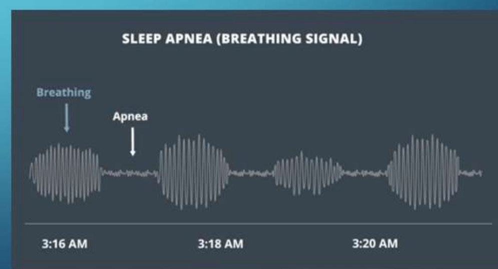
Image and descriptions by Dina Katabi, Andrew and Erna Viterbi Professor of EE and Computer Science, MIT

52

SLEEP

- Sleep studies are nonrepresentative
  - Unfamiliar place, bed, one night of sleep, wearing electrodes
- Emerald can measure sleep as well as the gold standard, but in situ
  - Awake
  - Light sleep
  - Deep sleep
  - REM
  - Sleep apnea

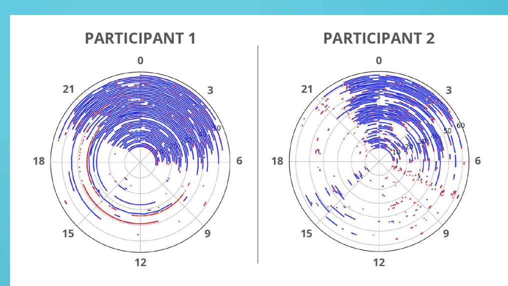


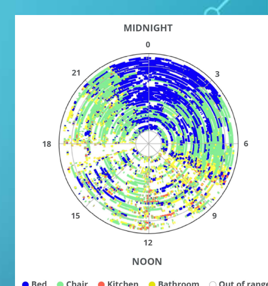


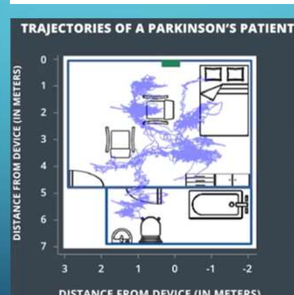
53

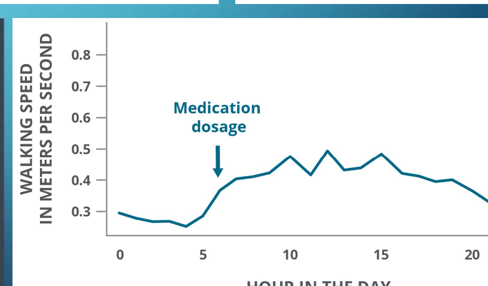
MOBILITY

- Mobility and associated patterns of health

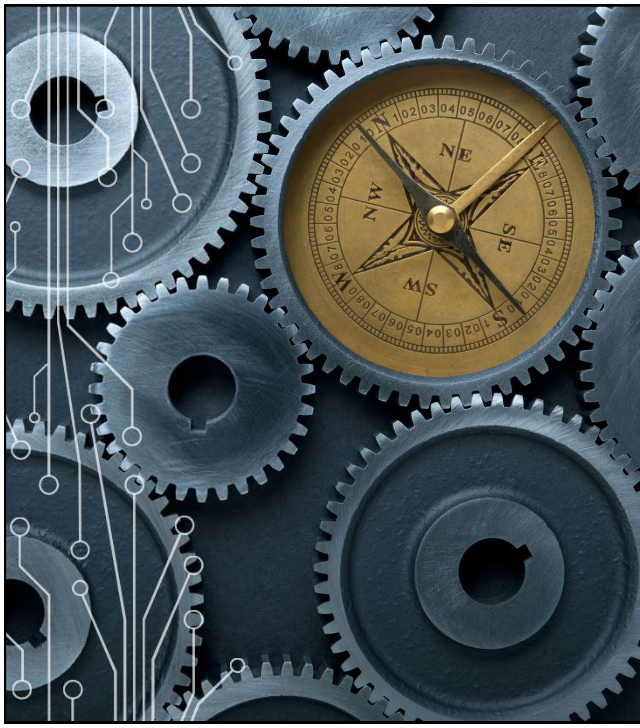








54



WHY THE FUTURE OF AI IS NOT QUITE BRIGHT ENOUGH FOR SHADES

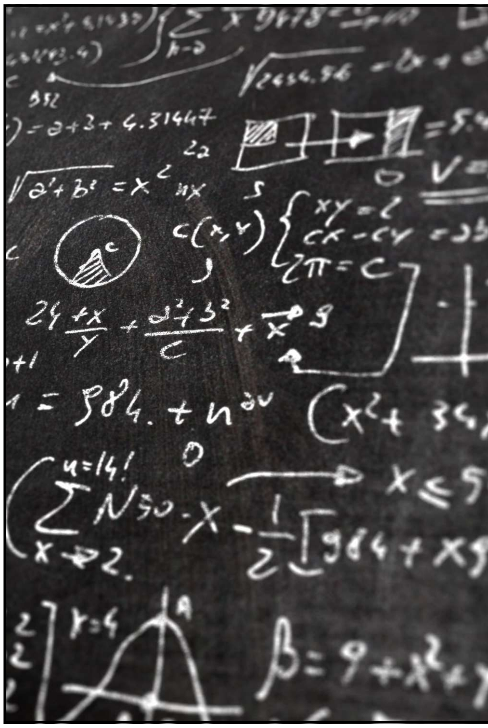
(WITH APOLOGIES TO TIMBUK3, AND THOSE TOO YOUNG TO KNOW WHO THEY WERE)

55

EXPENSIVE AND TIME-CONSUMING

- Annotation is expensive
  - Time of experts to classify cases and identify and label features
- Curation takes time and is easy to get wrong
  - Who makes up your population?
  - Different imaging machines, preprocessing data (typos, IDC-9 vs IDC-10)
- AI needs to be connected to the EHR to be most effective
  - Requires significant expertise
  - If you don't have expertise, have to hire outside help
  - AND must deidentify EHR data in ADDITION
- Complex problems require complex AI
  - Will need outside help in most cases
  - Interpretability will be challenging

56



UNDERSTANDING AND TRUST

- AI as a "Black Box"
- Power of AI: do things humans cannot do
  - Or do them faster
  - Often makes it less understandable
- Working to build interpretable AI
  - List the features to which it is "paying attention"
- Challenges of interpretability
  - Interpretable for doctor ≠ interpretable for patient
  - Interpretable at what level of detail?
    - Biological prediction of drug target?
    - Full biological mechanism of processing by body?

57

BUILDING INTERPRETABLE AI

- Test the model against humans
  - Build AI models that first explain, then provide outcome
  - Have panel of experts work through each case
- Adversarial training
  - Change cases in ways that should change the outcome and seeing how the AI output changes
- Ask AI to explain, have human make the final call
  - Learned intermediary
  - FDA often requires human as the final "decider"
  - Good model for AI

58

HOW AI CAN GO WRONG EVEN WHEN DOING OUR BEST

- Risk stratification and hospital optimization
  - Predict sickest patients and align resources and care
- Optum AI developed for use by hospitals
  - Data vectors included IDC codes
  - More codes = more treatment = sicker patient
- Results
  - White people tended to be classified as higher risk
  - Black people tended to be classified as lower risk
- Why?
  - IDC codes confounded with insurance levels and cost
  - Better insurance = more treatment ≠ sicker people!
- Explainability
  - If AI said what it was "paying attention" to…

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. doi:10.1126/science.aax2342

59

IMPORTANCE OF CHALLENGING AND REFINING MODELS

- Regina Barzilay (MIT) and Connie Lehman (Massachusetts General)
- During development of radiography AI for cancer risk stratification
- Selected images from patients with poor or good outcomes
- AI hit 99.9% AUC for images of patients with and without cancer
  - Selected all cancer cases from one year and all non-cancer from the next year
  - Mass Gen had switched imaging machines between those years
  - Each machine has unique stamps
- AI found the best, most predictive pattern in could
  - Machine type, not cancer

60

WHEN SUPERVISION CEASES: MENTAL HEALTH

- Tessa (shared via National Eating Disorders Association)
- Initially a closed system using SML (NLP)
- Later bought by company that connected to generative AI (unsupervised)
- Started giving incorrect advice under repetitive questioning

61

GENERALIZABILITY IN AI

- When you've seen one deep learning model….
  - You've seen one deep learning model
- Predicting patient outcomes of medication for schizophrenia
  - Deep learning performed very well for members of the training data set
  - No better than chance when applied to other datasets
    - Chekroud AM, Hawrilenko M, Hieronimus Loho, et al. Illusory generalizability of clinical prediction models. Science. 2024;383(6679):164-167. DOI: 10.1126/science.adg8538

62

THE (NEAR) FUTURE OF AI

- AI factories
  - Automates data collection, preprocessing, labeling, augmentation
  - Continually integrates and adjusts to incoming (new) data
  - Preliminary model development
  - Selects and “tunes” algorithms to refine and select best models
  - Deploys testing frameworks to test models against validation datasets
  - Monitors model performance, alerts to model drift and anomalies
- Will exponentially accelerate deep learning development and remove expertise and resource barriers
- Must be used by humans as productivity tool for SML, not as automomous stand-alone

63

CLOSING THOUGHTS

- Future of AI in health care is already here
- Today’s AI is the worst you will ever use
- In some cases, already better than humans
- Not “Yes or No,” but when, how, for whom and for what problems?

64

WORKS CITED AND ADDITIONAL READINGS

- Alvarez-Melis, David and Tommi S. Jaakkola. "A Causal Framework for Explaining the Predictions of Black-Box Sequence-to-Sequence Models." ArXiv:1707.01943, November 14, 2017. http://arxiv.org/abs/1707.01943.
- Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. "Synthesizing Robust Adversarial Examples." Last modified June 7, 2018. http://arxiv.org/abs/1707.07397.
- Bahl, Manisha, Regina Barzilay, Adam B. Yedidia, Nicholas J. Locascio, Lili Yu, and Constance D. Lehman. 2018. "High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision." Radiology 286, no. 3: 810–818. https://doi.org/10.1148/radiol.2017170549.
- Bennett, Charles L., Tammy J. Stinson, Victor Vogel, Lyn Robertson, Donald Leedy, Patrick O'Brien, Jane Hobbs, et al. 2000. "Evaluating the Financial Impact of Clinical Trials in Oncology: Results From a Pilot Study From the Association of American Cancer Institutes/Northwestern University Clinical Trials Costs and Charges Project." Journal of Clinical Oncology 18, no. 15: 2805–2810. https://doi.org/10.1200/JCO.2000.18.15.2805.
- Bertsimas, Dimitris, and Jean Pauphilet. Forthcoming. "Holistic Hospital Optimization." Management Science.
- Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. "Domain-Adversarial Training of Neural Networks." Last modified May 26, 2016. https://arxiv.org/abs/1505.07818.
- Hartmann, Lynn C., Thomas A. Sellers, Marlene H. Frost, Wilma L. Lingle, Amy C. Degnim, Karthik Ghosh, and Robert A. Vierkant, et al. 2005. "Benign Breast Disease and the Risk of Breast Cancer." The New England Journal of Medicine 353: 229–237. https://www.nejm.org/doi/full/10.1056/NEJMoa044383.
- Kabelac, Zachary, Christopher G. Tarolli, Christopher Snyder, Blake Feldman, Alistair Glidden, Chen-Yu Hsu, Rumen Hristov, E. Ray Dorsey, and Dina Katabi. "Passive Monitoring at Home: A Pilot Study in Parkinson Disease." Digital Biomarkers 3, no. 1 (April 30, 2019): 22–30. https://doi.org/10.1159/000498922.
- Khullar, Dhruv. 2019. "A.I. Could Worsen Health Disparities." New York Times, January 31, 2019. https://www.nytimes.com/2019/01/31/opinion/ai-bias-healthcare.html.
- Shaaban, A.M., J.P. Sloane, C.R. West, F.R. Moore, C. Jarvis, E.M. Williams, and C.S. Foster. 2002. "Histopathologic Types of Benign Breast Lesions and the Risk of Breast Cancer: Case-Control Study." The American Journal of Surgical Pathology 26, no. 4: 421–430. https://www.ncbi.nlm.nih.gov/pubmed/11914619.
- Yala, Adam, Regina Barzilay, Laura Salama, Molly Griffin, Grace Sollender, Aditya Bardia, Constance Lehman, et al. 2017. "Using Machine Learning to Parse Breast Pathology Reports." Breast Cancer Research and Treatment 161: 203–211. https://doi.org/10.1007/s10549-016-4035-1.

65

QUESTIONS?

66