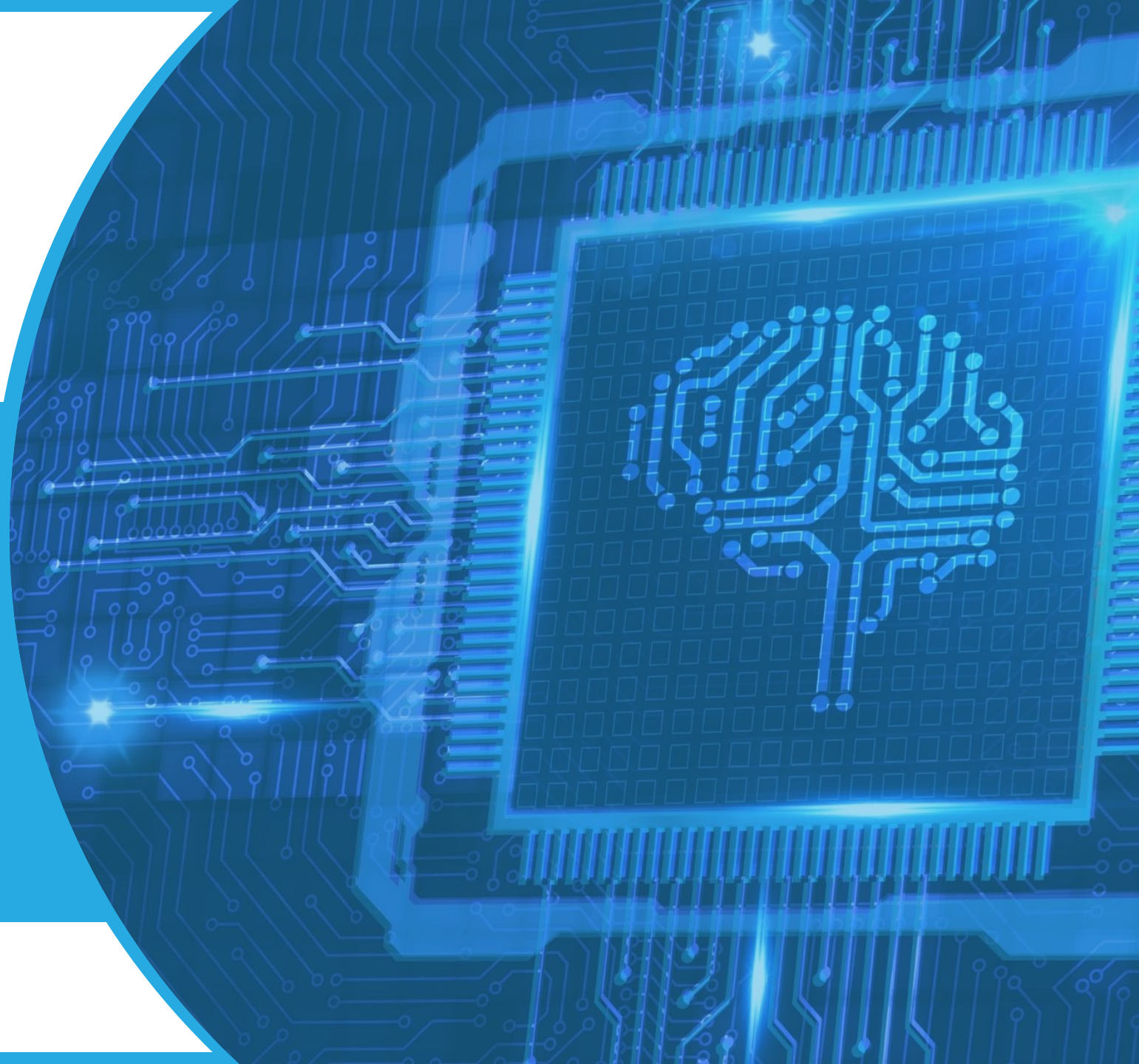
ISPOR

Improving healthcare decisions

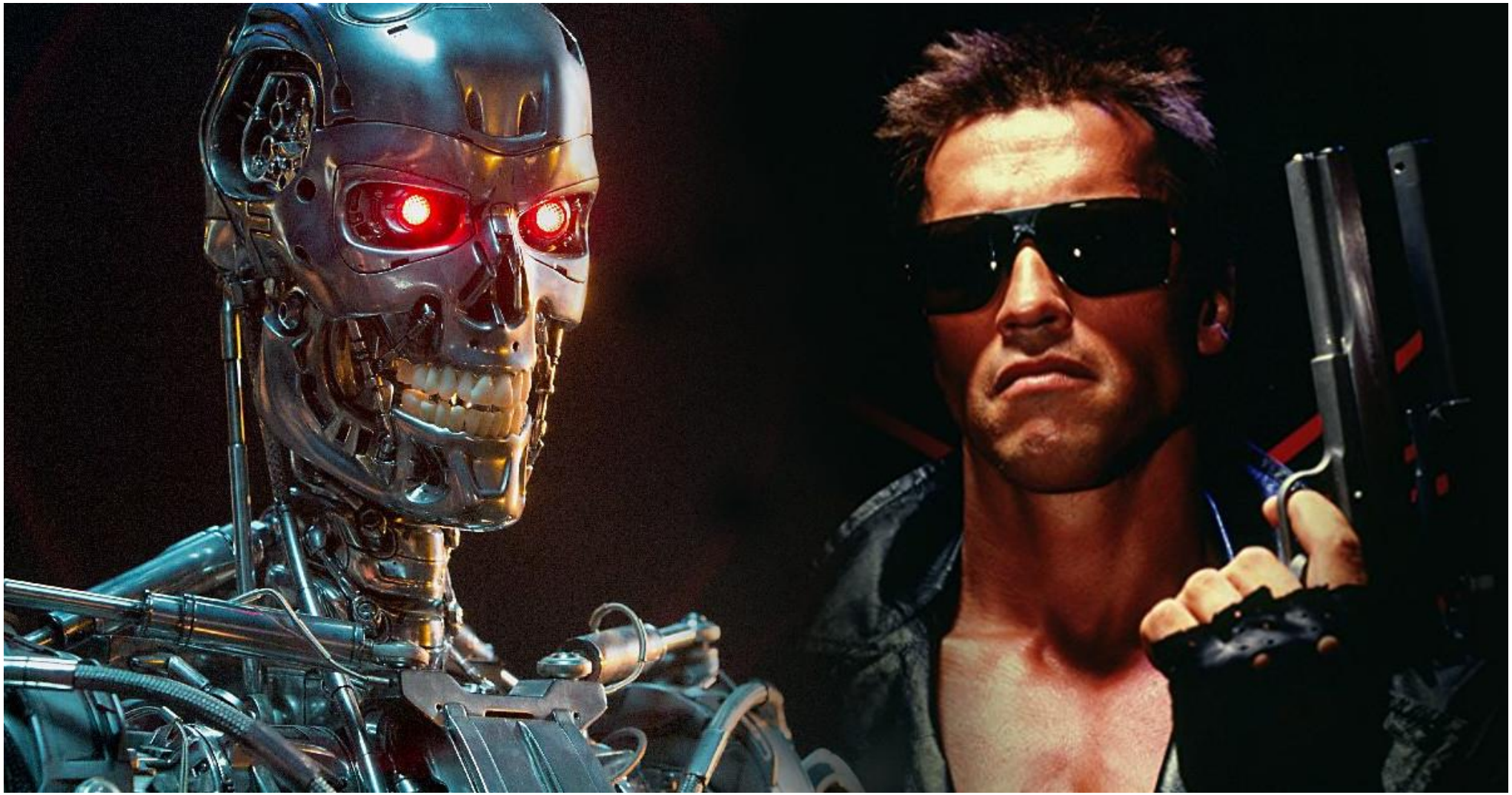
Making Machine Learning a Reality for HEOR - Practical Applications to Real-World Studies

Wei-Shi (Danny) Yeh, PhD
Genentech
South San Francisco, CA, USA

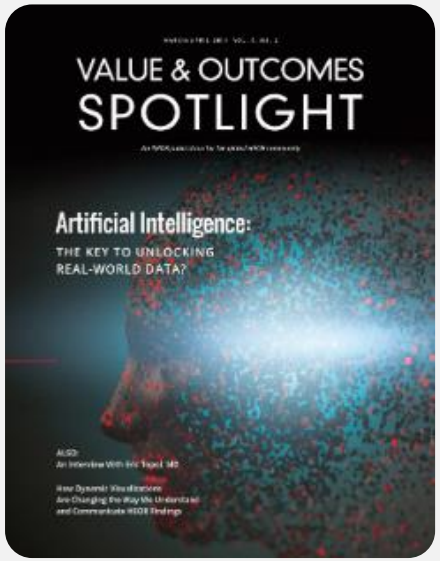
May 20, 2020

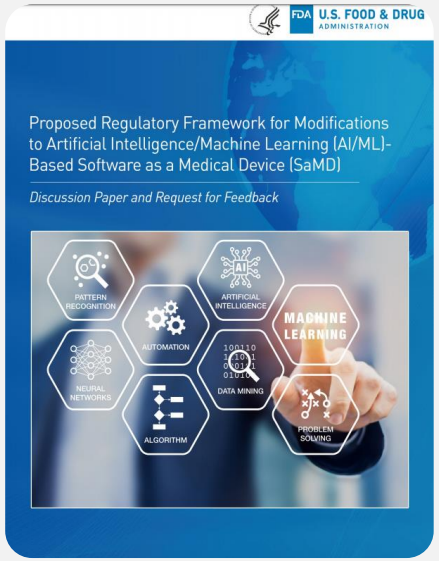


The First Time I Heard about Artificial Intelligence (AI)



Artificial Intelligence (AI) and Machine Learning (ML) are Trendy Topics in Healthcare





- **Excitement about the future applications**
- **Vendor capability presentations**
- **Candidate capabilities**

Sources:
https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
https://onlinedegrees.sandiego.edu/big-tech-invest-health-care-innovations-informatics/
https://www.ispor.org/publications/journals/value-outcomes-spotlight/issue/march-april-2019

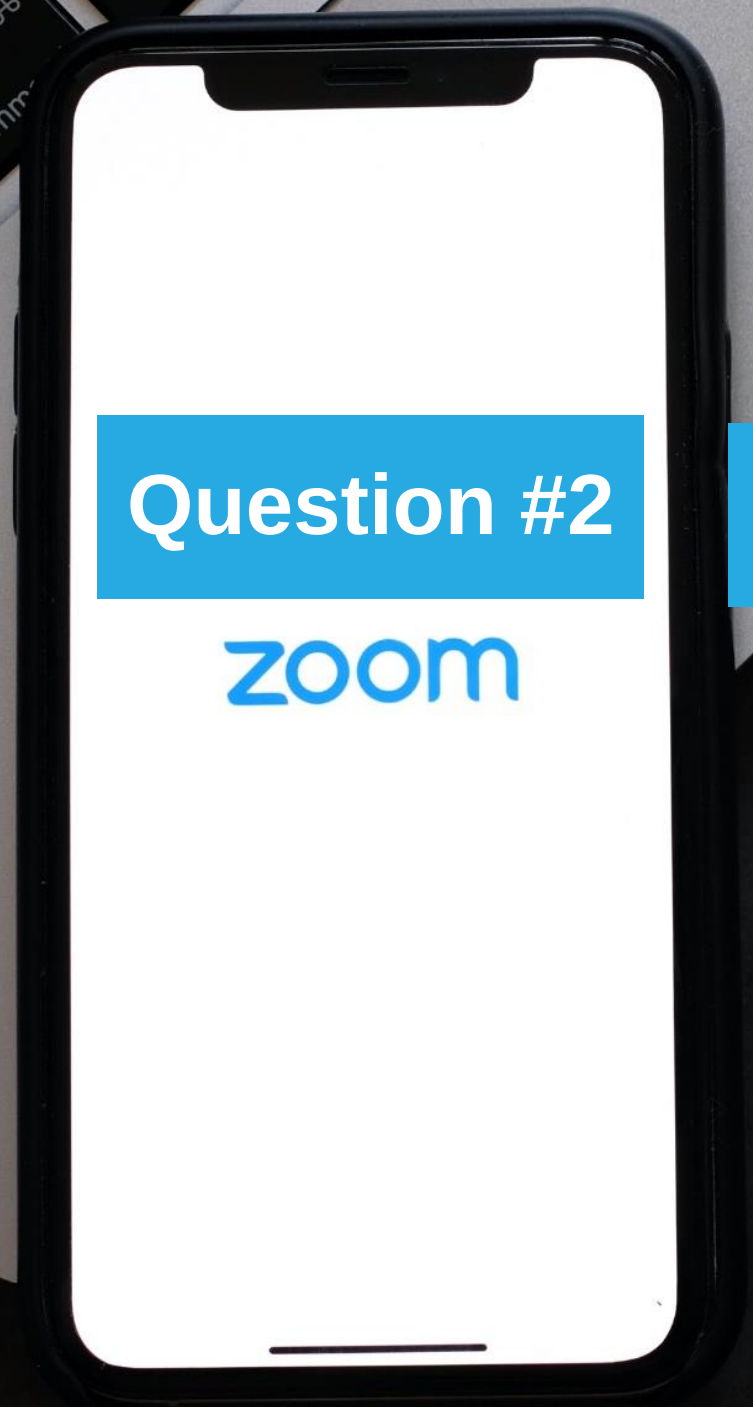
Question #1

Have you used AI/ML in your research?

1 Yes

2 No

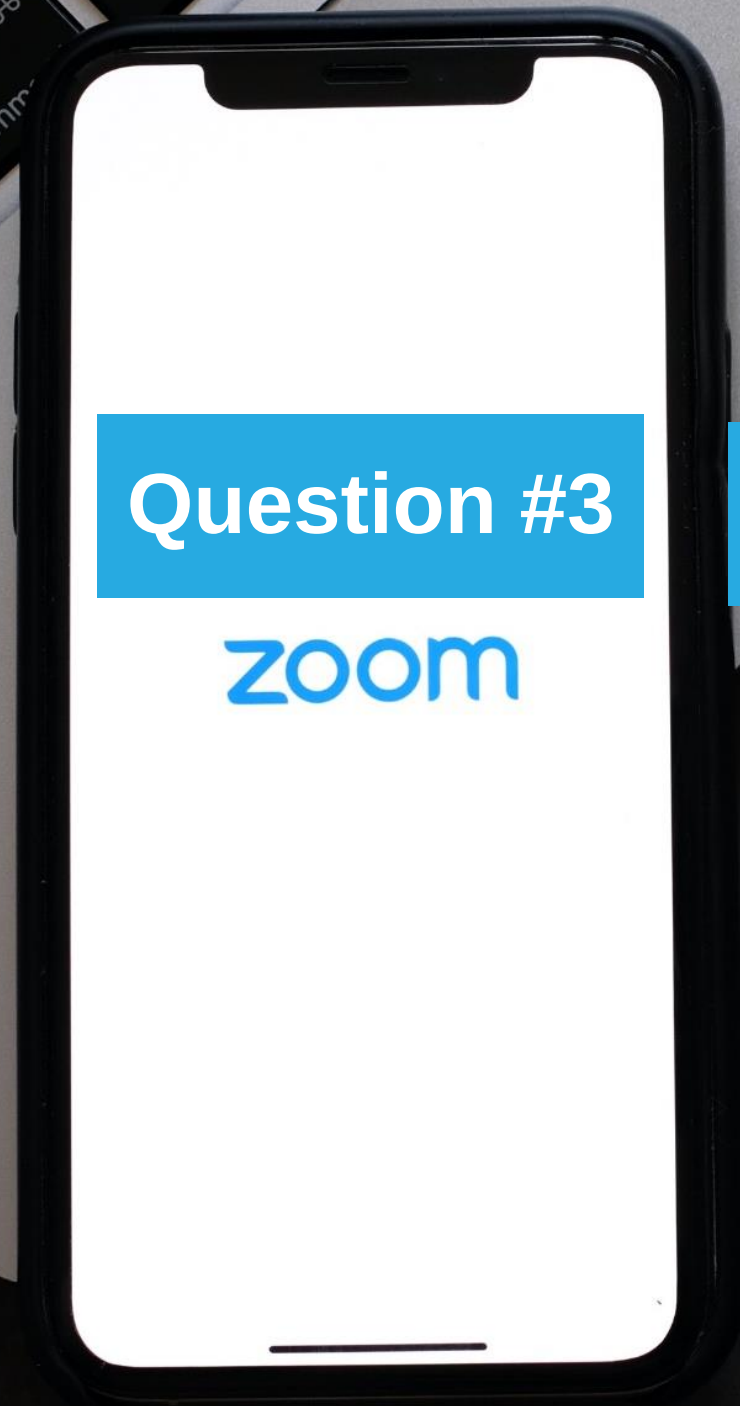
Question #2



Do you expect to use AI/ML in your own research in the next 5 years?

1. Yes
2. No

Question #3



Compared to 2020, how frequent will AI/ML be used in the HEOR research in 2025?

1. Definitely more
2. Slightly more
3. The Same
4. Probably less

Opportunities for AI/ML (1) Reduce Human Efforts for Tedious Evidence Synthesis

Extracting unstructured information into a structured format quickly

- Scientific literature
- Patient chart
- Conversation transcript
- Website information
- Medical image

Opportunities for AI/ML (2) Enhancing Predictability through Hypothesis Free Analyses

Identify additional Xs that matter to Y

- Symptoms/comorbidities
- Procedures
- Interactions

Enhance predictability

- Diagnosis
- Cost/resource utilization
- Clinical outcomes

Opportunities for AI/ML (3)
Combining the Strengths of Both

Extracting information

Hypothesis free analysis

Better RWE Research

Applications in RWE Research

Diagnosis

- Refine diagnostic algorithms
- Identify potential patients

Outcomes

- Refine regression models
- Identify safety signals

Treatment pattern

- Identify signals

Barriers for AI/ML to be Widely Adapted

Hypothesis free approach vs. pre-specified analysis plan

Reproducibility by others

Validation vs. traditional approach

Panelist

Agustin Marquez, MS

nference, Vice President of Product, BD and Marketing

Leveraging AI to "curate" at scale unstructured data from biomedical literature and medical records

Danny Sheinson, PhD

Genentech, Principal Data Scientist, US Medical Affairs

Applying insights from AI/ML to HEOR studies

Jenny Lo Ciganic PhD

University of Florida, Assistant Professor, Pharmaceutical Outcomes and Policy

ML and prediction in HEOR research

Q&A session

www.ispor.org

ISPOR

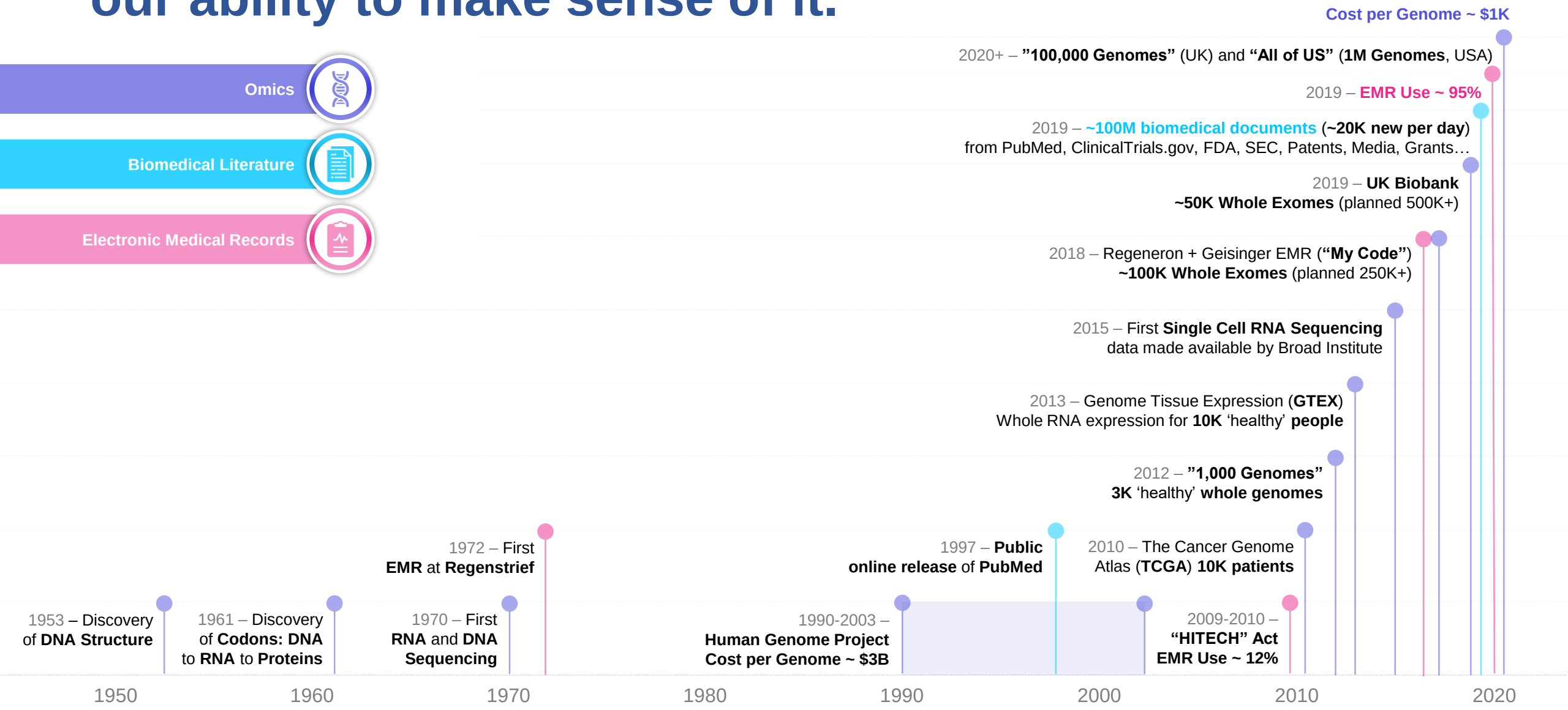
Improving healthcare decisions

Leveraging AI to “curate” at scale unstructured data from biomedical literature and medical records

Agustin J. Lopez Marquez
nference
South San Francisco, CA, USA

May 20, 2020
ISPOR 2020, Virtual Conference

Biomedical data is exploding… at a pace that exceeds our ability to make sense of it.



Humans can read only a tiny fraction of all biomedical knowledge. Machines can help.

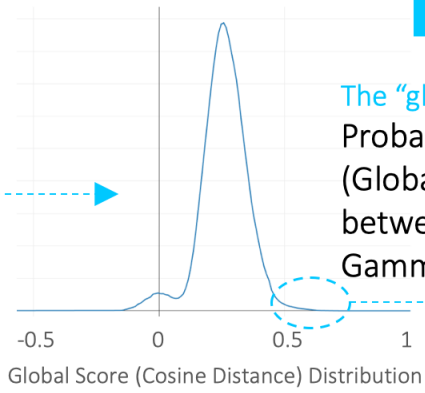
Unstructured Text Data

900B+ words in ~100M docs
(200M+ unique vectors)

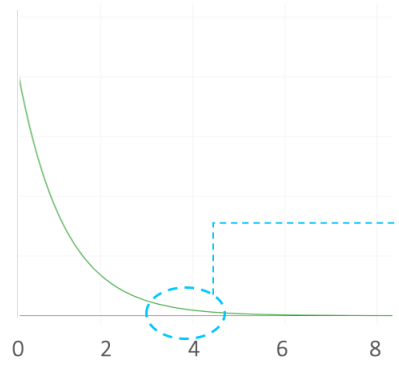
nferX
Neural Networks

Word Embedding Network (e.g. word2vec) and Autoencoders
BiLSTM Network (Bidirectional Long Short Term Memory)
Convolutional Neural Network
Reinforcement Learning Network

Statistical Inference



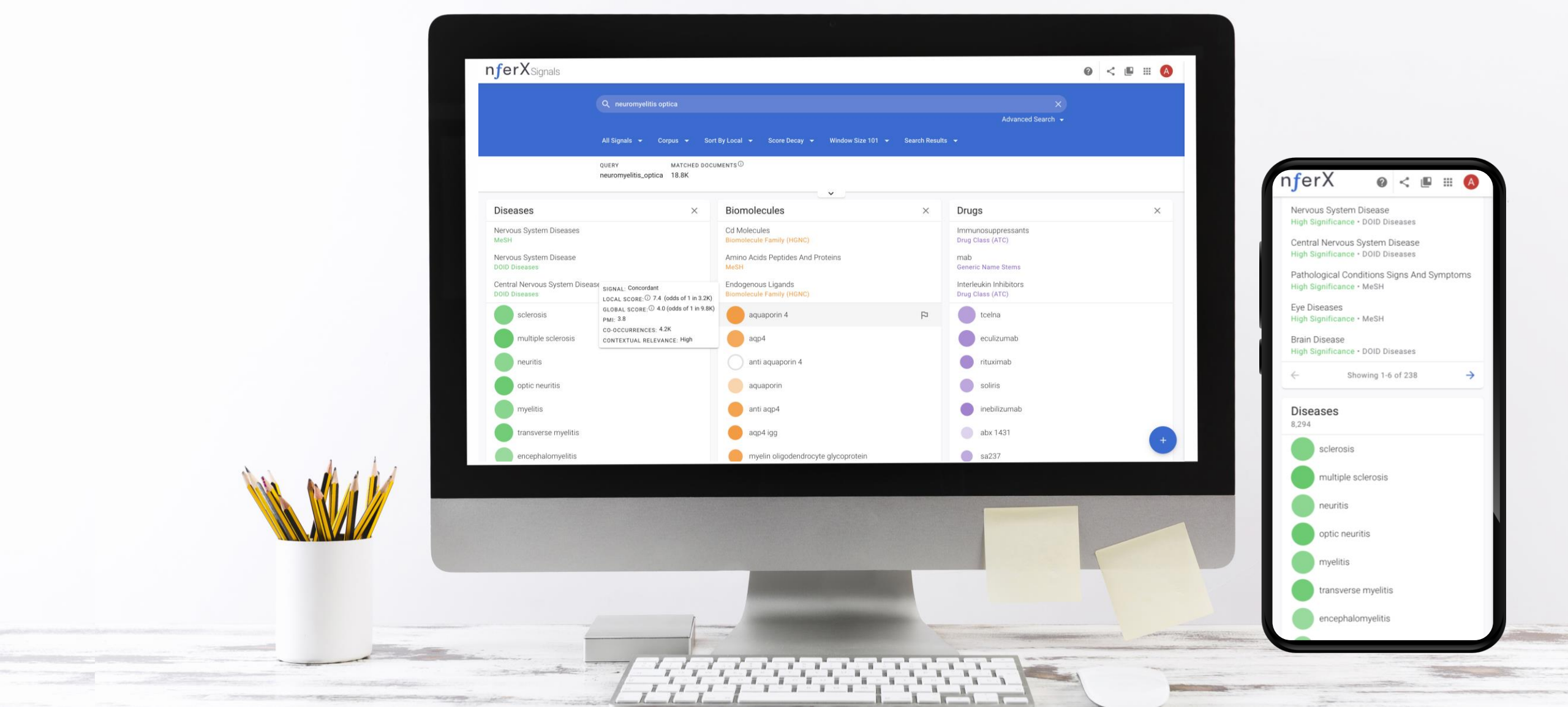
The "global" approach leverages AI (deep learning)
Probability distribution of the cosine distance (Global Score before conversion to neglogp scale) between pair of vectors in Output:
Gamma Mixture distribution



The "local" approach leverages traditional machine learning (natural language processing)
Probability distribution of the Local Context Score between pairs in Output: Exponential distribution

Statistically significant associations between phrases

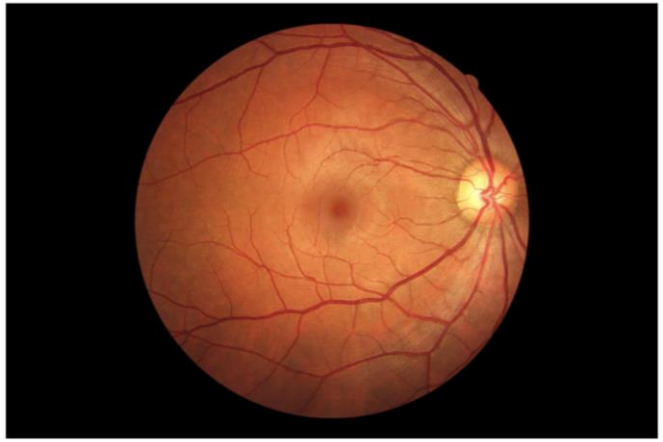
How can machines help? NLP can help us “curate” unstructured text information at scale.

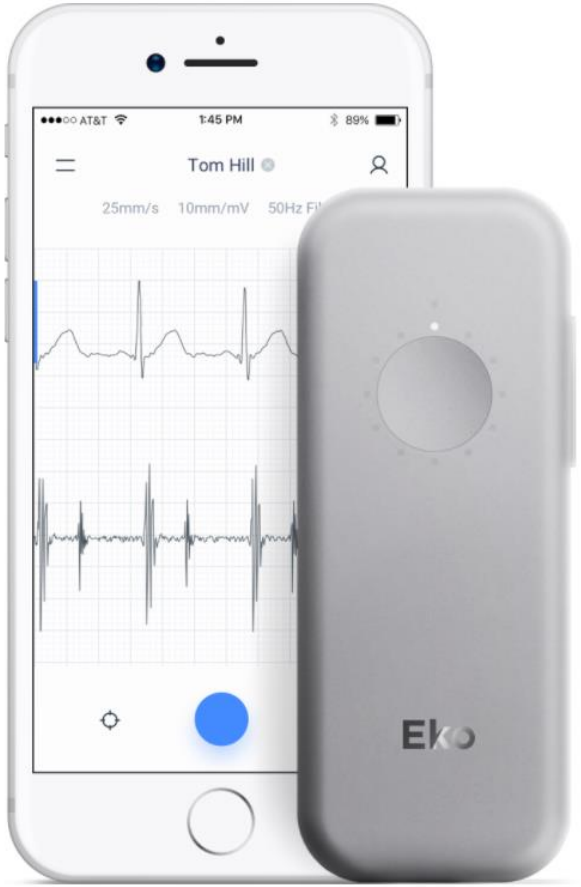


AI models are showing a lot of promise in medicine… but…

It's All in the Eyes: Google AI Calculates Cardiovascular Risk From Retinal Images

Synced Follow
Feb 20, 2018 · 3 min read





Mayo Clinic Uses AI to Glean Patients' Overall Health From EKG Heart Test

AI may be able to help find diseases in patients, Mayo researcher says

THE WALL STREET JOURNAL. August 29, 2019

"Because AI has profound capabilities to detect subtle patterns in biomedical data, it may help us find—and it looks like it can help us find—disease people have that are still occult," said Paul Friedman, a member of the research team and chair of the Department of Cardiovascular Medicine at Mayo, a medical center in Rochester, Minn.

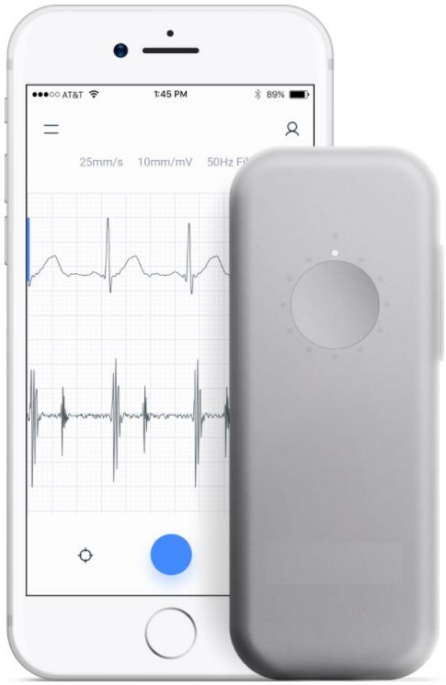
Other members of the medical community are excited about the possibilities Mayo is uncovering.

"They are just extracting things about the 12-lead cardiogram that, I have to say, would be unimaginable without this use of this deep-learning capability," said Eric Topol, a cardiologist and executive vice president at Scripps Research.

Paul Friedman, MD
Chair of Cardio at Mayo
(#1 Cardio Group in US)

nference Collaborator at Mayo Clinic for Pulmonary Hypertension Project

… manual curation is often a bottleneck to create “labeled data” that can train AI models.



Digitized Images

Current Approach

Major Pain Point/Bottleneck

Manual curation of important patient characteristics (such as presence/absence of disease, treatments, outcomes, etc.) from unstructured clinical records by residents, nurses, etc.

FAMILY HISTORY:

Her father is alive with history of psoriasis and hypertension. Mother is deceased. She did have scleroderma. She does have one brother alive, and one brother passed around childbirth with a heart condition. She does have two sisters alive and two sons. No family history of multiple myeloma.

HISTORY OF PRESENT ILLNESS:

Mrs. Name presented to her local emergency room in Date with back pain. They did x-rays done initially at that time, and they felt that this was probably muscle spasms, and she was treated with muscle relaxants. The back pain however continued to worsen and was seen in the emergency room subsequently on two different occasions, and finally had an MRI of the spine done which showed a T8 vertebral compression fracture or possible hemangioma. She was admitted to Location where discussion was had at that time as to what further therapeutic intervention she needed. She however decided to come to Organization Location for a second opinion before proceeding.

She was seen first here on Date by Dr. Name in Orthopedics. At that time, he did not feel she needed to have a vertebroplasty done immediately and recommended conservative treatment with a brace. She returned for followup in Date and at that time based on the MRI findings, the decision was made to proceed to a CT-guided biopsy of the T8 vertebral body lesion.

Preliminary results suggest a plasma cell neoplasm, but the final results are still pending. In looking back,she has also noticed some intermittent rib pain. Outside of this, she has not any other symptoms. She denies any fever, chills, shortness of breath, palpitations, or paroxysmal nocturnal dyspnea. Not any change in energy level.

ECOG performance status of 0. Name of 90%.

Labeled Patient-Level Data

AI/Machine Learning Models

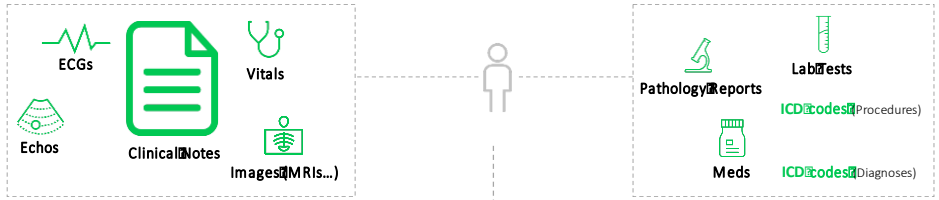
Why is “curation” of clinical records important? Structured data – such as ICD codes – often provide an inaccurate and/or incomplete picture of the real patient’s clinical information.

flatiron

Example: Advanced NSCLC

Variable	Structured data only	Flatiron data completeness
Metastatic diagnosis	26%	100%
Smoking status	0%[1]	94%
Histology	37%	99%[2]
Stage	61%	95%
ALK results (of those tested)	9%	100%[3]

nference's "augmented curation" – leveraging AI - dramatically improves the speed and scale of curation with the same accuracy as manual curation (current gold standard).



Curation: Status Quo

flatiron

90%+ data from community centers

Oncology | Immunology | Cardio & Metabolism | Rare Diseases | Infectious Diseases | Neuroscience

~2M patients (raw data)

MANUAL CURATION (BOTTLENECK)

~1,000 abstractors (predominantly nurses…)

~0.2M patients (curated data)

nference MAYO CLINIC

67% data from **centers of excellence**

Oncology | Immunology | Cardio & Metabolism | Rare Diseases | Infectious Diseases | Neuroscience

~9.8M patients (raw data)

AUGMENTED CURATION

~10's scientists (MD/PhDs from Top Institutions) perform **manual curation** to **train state-of-the-art neural networks**

~9.8M patients (curated data)

Let’s see how “augmented curation” can help in a problem that matters: PAH early diagnosis.

Early detection and management of pulmonary arterial hypertension

Marc Humbert, J. Gerry Coghlan, Dinesh Khanna

European Respiratory Review 2012 21: 306-312; **DOI:** 10.1183/09059180.00005112

Article | Figures & Data | Info & Metrics | PDF

Abstract

The long-term prognosis for patients with pulmonary arterial hypertension (PAH) remains poor, despite advances in treatment options that have been made in the past few decades. Recent evidence suggests that World Health Organization functional class I or II patients have significantly better long-term survival rates than patients in higher functional classes, thus providing a rationale for earlier diagnosis and treatment of PAH. However, early diagnosis is challenging and there is frequently a delay between symptom onset and diagnosis. Screening programmes play an important role in PAH detection and expert opinion favours echocardiographic screening of asymptomatic patients who may be predisposed to the development of PAH (*i.e.* those with systemic sclerosis or sickle cell disease), although current guidelines only recommend annual echocardiographic screening in symptomatic patients. This article reviews the currently available screening programmes, including their limitations, and describes alternative screening approaches that may identify more effectively those patients who require right heart catheterisation for a definitive PAH diagnosis.

Mayo and nference are building AI models to predict early-presentation of pulmonary hypertension using ECG images. However, the biggest bottleneck is: manual curation.

1,630

Pulmonary hypertension patients that were “manually” confirmed as “positive controls” by residents/nurses.

The residents/nurses had to go through every record in the Pulmonary Hypertension clinic at Mayo Clinic to determine from the clinical notes whether a patient had pulmonary hypertension.

~10%

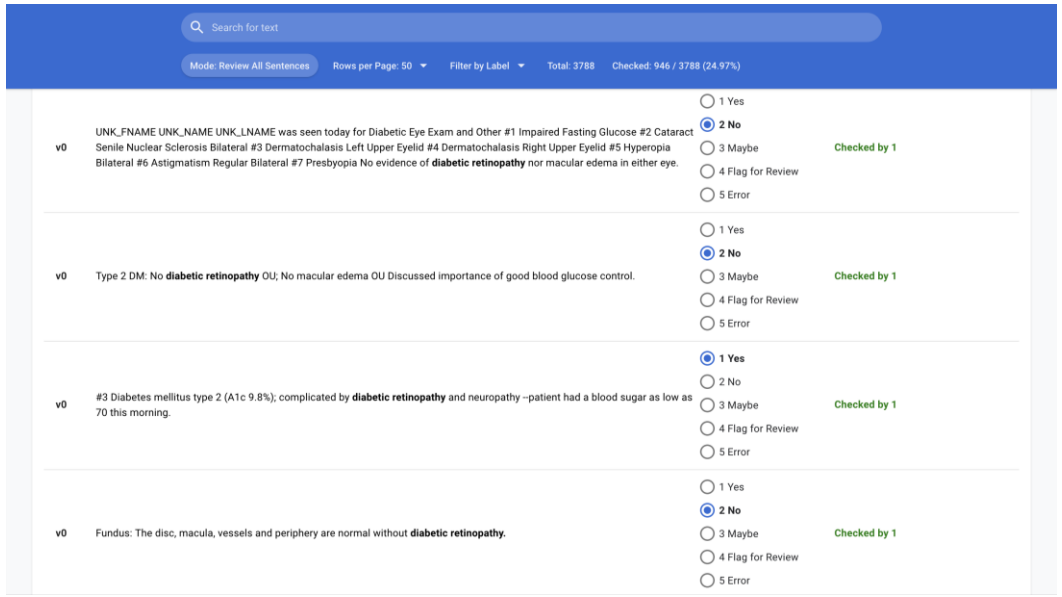
After months of “manual curation”, the nurses/residents have reviewed only approx. 10% of all possible pulmonary hypertension records to identify “positive” and “negative” controls due to how labor intensive this process is.

nference's solution: "augmented curation" leveraging state-of-the-art NLP (BERT models)

How does "augmented curation" work?

Step 1: nference developed a software ("BYOB"*) that allows our scientists to tag sentences (coming from the clinical notes) for a specific task.

In this particular example, the software was configured to show our scientists sentences – from the clinical notes at Mayo - that contain a given disease/phenotype (e.g. pulmonary hypertension). The scientists' job was to tag whether the patient had (or not) that particular phenotype from the context of each sentence.



Step 2: Our scientists tag a few thousands sentences for that specific problem and then train a neural network ("BERT"*) to automate this task.

Step 3: The neural network is tested – against manually-curated sentences – to measure its accuracy. If the accuracy is not sufficiently high, the scientists go back to tagging more sentences until the accuracy is the desired one. The power of this iterative approach is that the time spent in tagging after each new iteration is a lot shorter as the neural network becomes smarter.

iterative process

* BYOB stands for "Build Your Own BERT" where BERT stands for Bidirectional Encoder Representations from Transformers… which is a class of state-of-the-art neural networks for natural language processing.

nference's "augmented curation" outperforms "manual curation" (current gold standard)

1,630

Pulmonary hypertension patients that were "manually" confirmed as "positive controls" by Mayo Clinic residents/nurses

1,611 (98.8%)

Our algorithm correctly detected presence of pulmonary hypertension in the notes of these patients

19 (1.2%)

Our augmented curation model did not detect presence of pulmonary hypertension in 19 patients

Manual inspection of the notes showed that these cases had been incorrectly tagged by Mayo's staff (see example on the right)

(excerpts from more recent notes)

\nThe patient was referred to **Pulmonary Hypertension** by Dr. [redacted] from Allergy, because the patient was on sildenafil.

\nI recommended to Mr. B-PER REG that he should follow up with Dr. [redacted], as I do not see that we can be of significant help to him in the **Pulmonary Hypertension** Clinic.

I understand that he is having issues with his insurance company, in that they are unwilling to pay for the sildenafil without a formal diagnosis of **pulmonary hypertension**.

\nEssentially, Mr. B-PER REG does not have **pulmonary hypertension**, and the use of sildenafil in patients status post Fontan procedure with protein-losing enteropathy is an off-label use of the medication, but does have some physiological basis, as it is an attempt to decrease the right heart pressures.

That indicated that Mr. B-PER REG does not have **pulmonary hypertension**.

ISPOR

Improving healthcare decisions

Applying insights from AI/ML to HEOR studies

Danny Sheinson, PhD
Genentech
South San Francisco, CA, USA

May 20, 2020
ISPOR 2020, Virtual Conference

How can we integrate insights from an AI platform to RWE studies?

Is AI the next big thing in HEOR?

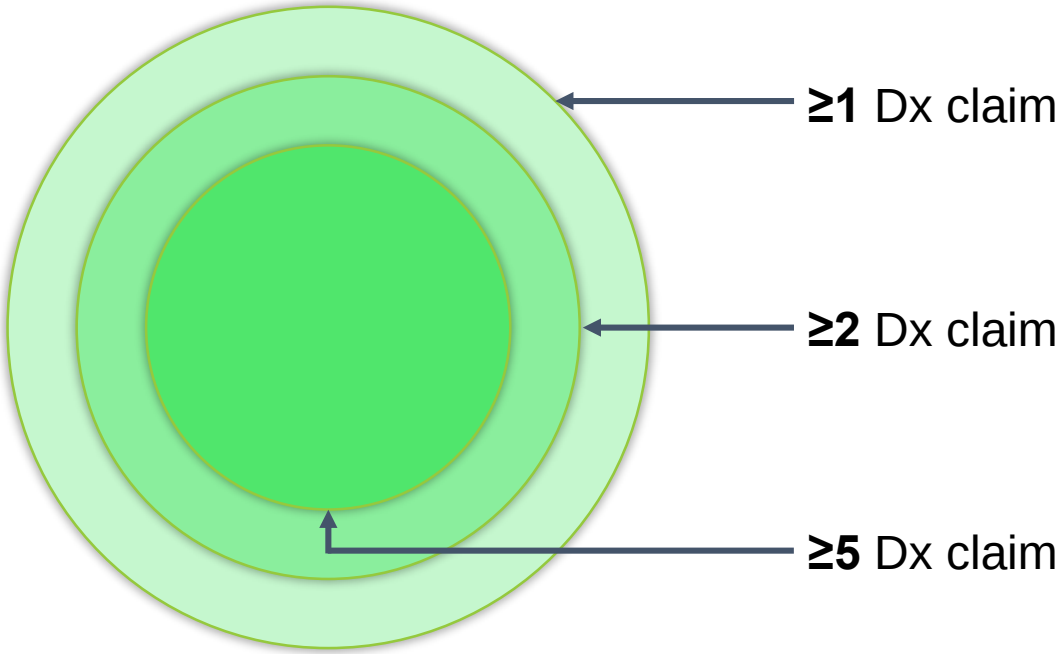
Two use cases

Improve identification of giant cell arteritis (GCA) patients in RWD

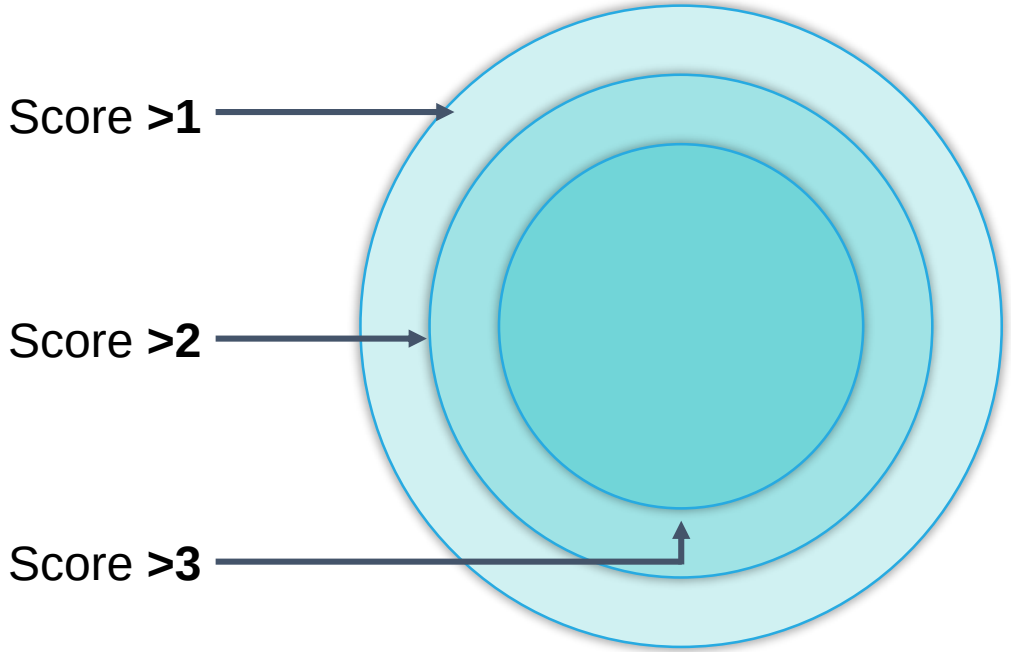
Identify risk factors associated with MS relapse

Can we leverage knowledge synthesized by nferX to improve the algorithm for identifying patients in claims data?

Traditional Approach

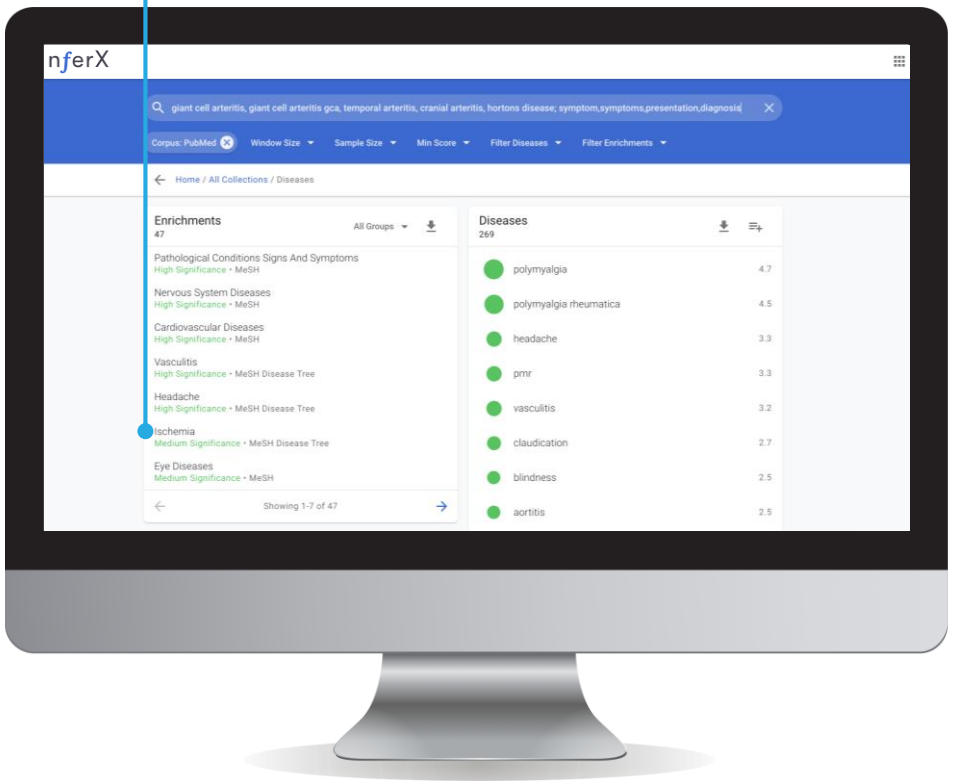


nferX Approach



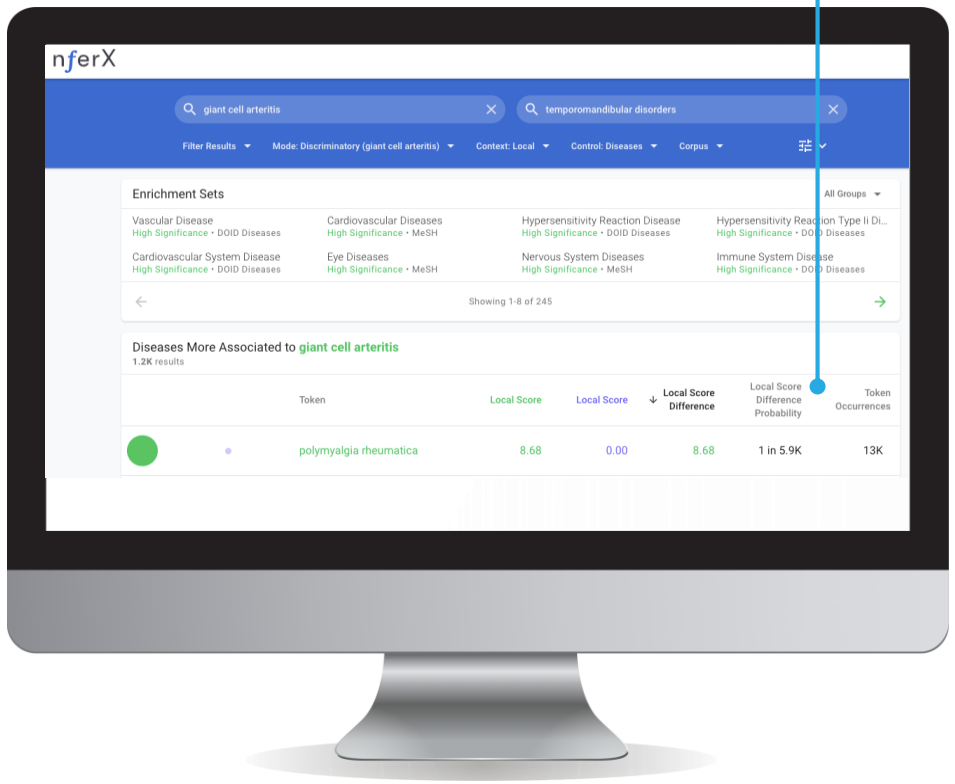
How can a comorbidity score measuring disease-specificity be derived from nferX?

nferX score quantifies the association to GCA



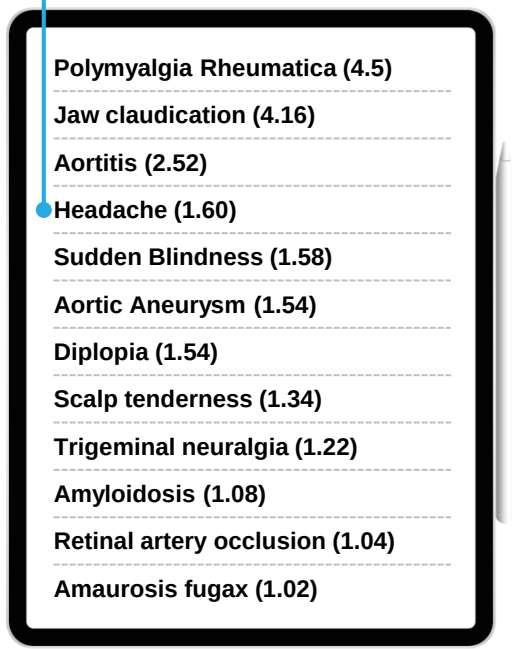
What are the disease phenotypes highly associated to GCA presentation and symptoms?

Polymyalgia rheumatica is highly specific to GCA



How specific are these phenotypes to GCA compared to other diseases?

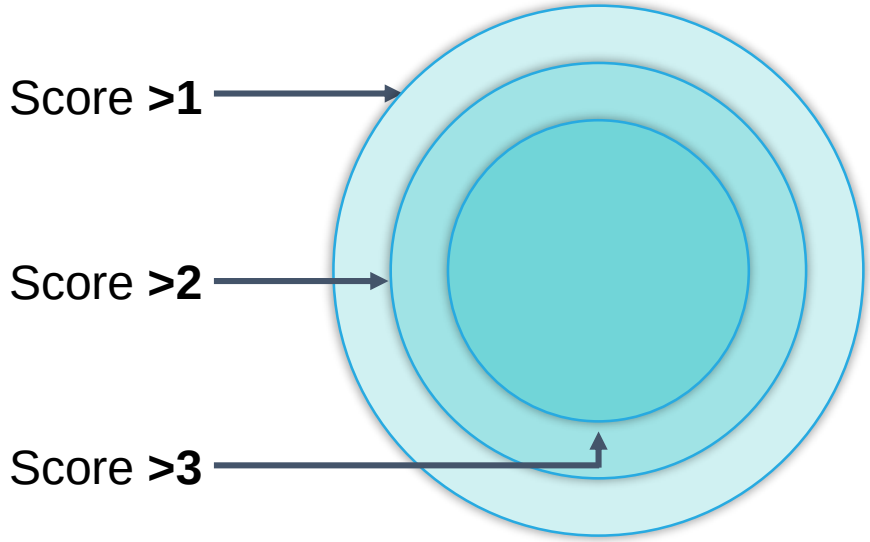
Composite weights can be used to calculate patient specific vectors



Weighted, GCA-specific phenotypes

Do comorbidity scores from nferX-derived weights correlate with typical patient selection steps in claims?

Cohort	N	Comorbidity Score
≥ 1 GCA claim between 2011 - 2015	2,945	0.99
≥ 1 inpatient claim or ≥ 2 outpatient claims between 2011 – 2015	2,056	1.97
≥ 5 GCA claim between 2011 - 2015	1,165	2.26



Next steps

What is the appropriate cutoff score?

- Predictive modeling / machine learning
- ISPOR 2020 Poster: Garcia-Rivera et al., *Identifying neuromyelitis optica patients from insurance claims data using nferX, a natural language processing-based platform*

How do we assess the performance? Validation against a gold standard

Common approach to model health outcomes and estimate treatment effects

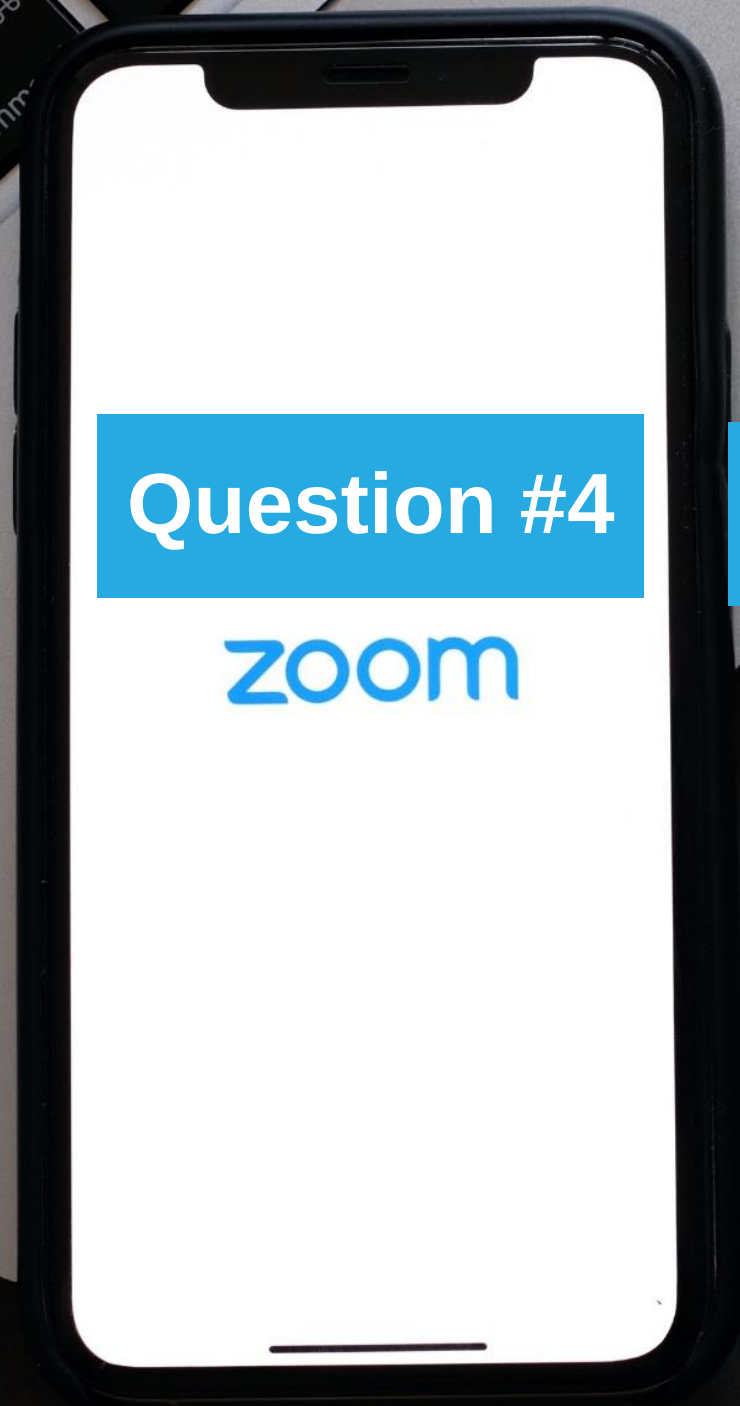
Illustrative example

Treatment

E (Health Outcome) $= f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \ldots + ?)$

Age, gender, race, other treatments, comorbidities, CCI, other important factors

Question #4



What is your most common approach to identifying adjustment covariates?

1. Review literature
2. Ask a clinician
3. Other

Common approach to model health outcomes and estimate treatment effects

Illustrative example

nferX

Treatment

E (Health Outcome) $= f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \ldots + ?)$

Age, gender, race, other treatments, comorbidities, CCI, other important factors

Example: Comparative effectiveness research of disease-modifying therapies (DMT) for the management of multiple sclerosis (MS)

Boster et. al, *Neurology and Therapy (2017)*

Neurol Ther (2017) 6:91–102
DOI 10.1007/s40120-017-0064-x

CrossMark

ORIGINAL RESEARCH

Comparative Effectiveness Research of Disease-Modifying Therapies for the Management of Multiple Sclerosis: Analysis of a Large Health Insurance Claims Database

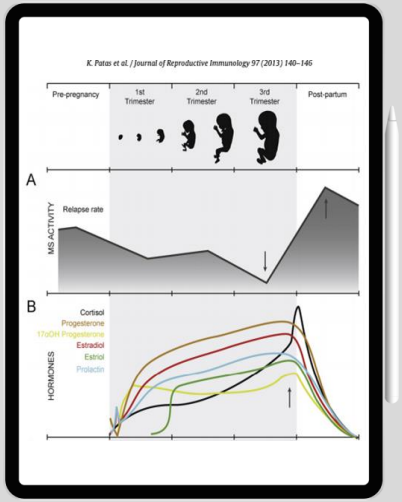
Aaron Boster · Jacqueline Nicholas · Ning Wu · Wei-Shi Yeh · Monica Fay · Michael Edwards · Ming-Yi Huang · Andrew Lee

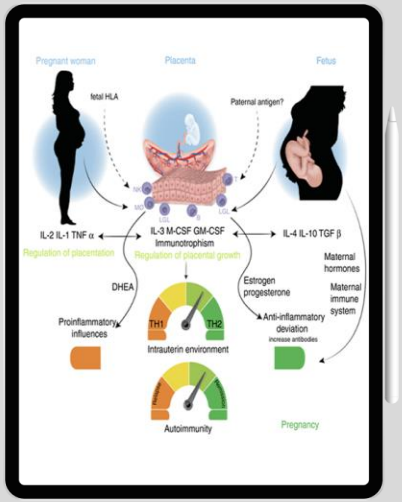
E(MS relapse rate 1-year post-index) = f(Treatments, patient demographics, CCI, pre-index relapse rate, use of other DMT, **individual MS-related symptoms**, other confounders)

nferX ?

MS-related symptoms, %

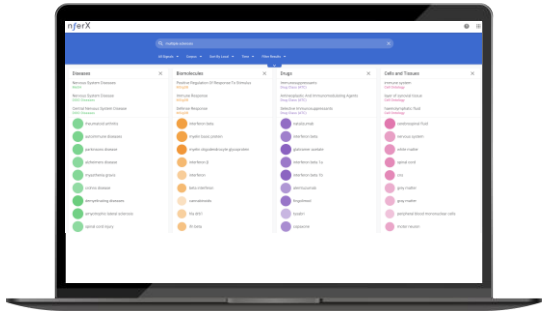
- Other causes of myelitis
- Demyelinating disease of CNS, unspecified
- Disorders of optic nerve and visual pathways
- Neurogenic bladder NOS
- Other disorders of soft tissues: neuralgia, neuritis, and radiculitis, unspecified
- General symptoms: dizziness and giddiness
- General symptoms: fatigue/malaise



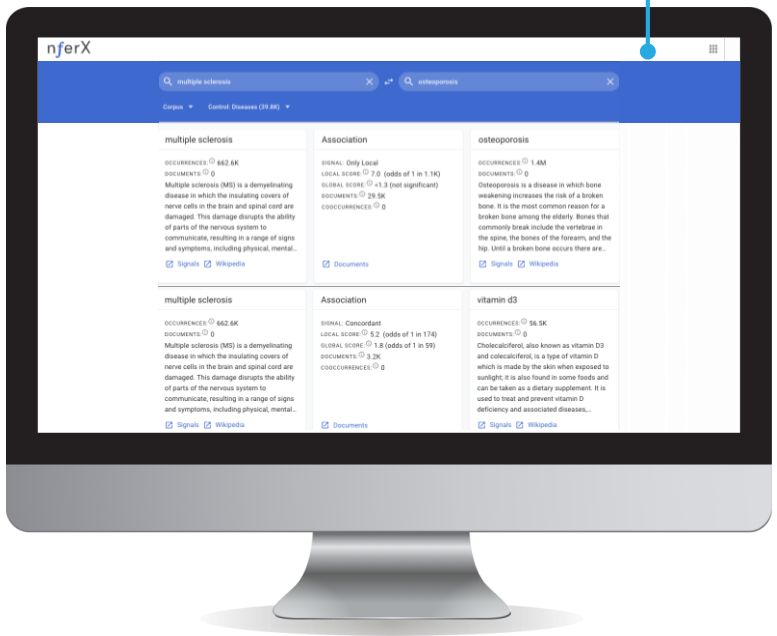


Relapse rates decline during **pregnancy** and increase in the **postpartum** period (VD level, hormone level, cytokines change)

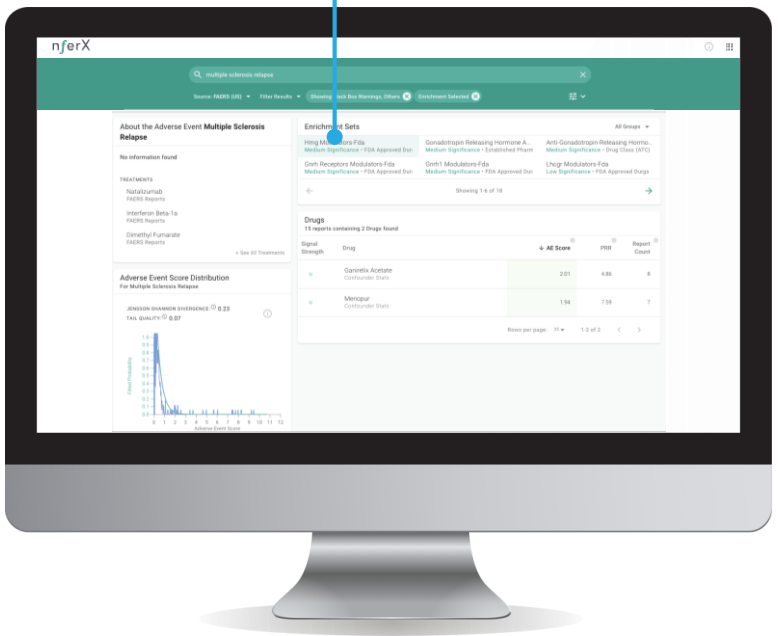
Several risk factors were identified via nferX



Osteoporosis / low vitamin D levels are associated with an increase in MS relapse rates



MS worsen due to withdrawal of estrogen during **menopause**

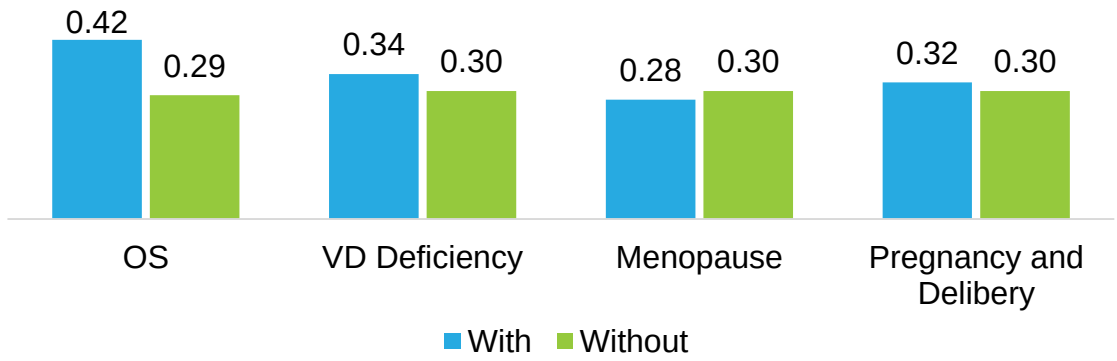


Revisit Boster (2017): Incorporating additional risk factors

Other factors	Disease-modifying treatments				
	D	I	G	T	F
N of patient	4568	849	1395	927	621
% of patients with the risk factors					
Pregnancy/child birth,%	1.9	4.4	4.7	0.5	2.6
Vitamin D deficiency,%	15.4	13.4	13.3	12.7	16.0
Osteoporosis,%	7.0	5.5	6.4	7.6	6.2
Menopause,%	3.2	2.6	2.0	4.2	2.8

*__D__= Dimethyl Fumarate | __I__= Interferon | __G__= Glatiramer | __T__= Teriflunomide | __F__= Fingolimod

MS relapse rate



Model parameter estimates of the risk factors are consistent with the descriptive

Estimated treatment effects remain largely the same

Model goodness-of-fit improved (lower AIC, significant LR test statistics)

ISPOR

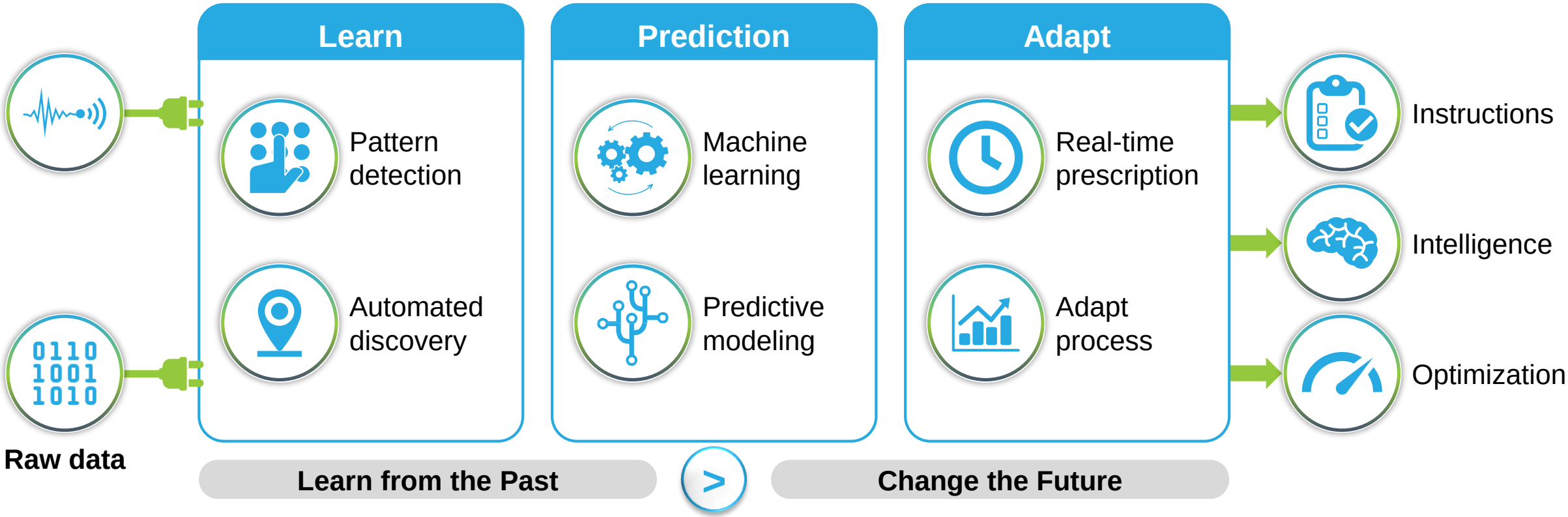
Improving healthcare decisions

Applying machine learning for prediction in health outcomes research

Jenny Lo-Ciganic, MSPharm, MS, PhD
Department of Pharmaceutical Outcomes & Policy
University of Florida, USA

May 20, 2020 | ISPOR 2020, Virtual Conference

Machine Learning & Prediction in a NutShell



Adapted from https://www.linkedin.com/pulse/smart-connected-service-operationalizing-machine-learning-ray-martin

Type of Machine Learning & Techniques

Supervised/Semi-supervised	Unsupervised	Reinforcement
Classification/categorization (e.g., death) • **Decision trees** • **Random forests** • **Forecasting** • **Support vector machines** • **LASSO** • **Elastic net** • **Bayesian statistics** • **Logistic regression** • Neural network **Regression analysis (e.g., cost)** • Regression trees • Penalized regression	• Clustering/trajectories • Affinity analysis (e.g., market basket analysis) • **K-mean clustering** • **Nearest-neighbor mapping** • Self-organizing maps • Singular value decomposition • Dimensionality reduction (e.g., image data)	• Artificial neural networks (ANN) • Learning automata • Markov decision process • Q-learning

Hot topics/techniques: **Deep learning (image diagnosis), natural language processing (e.g., sentiment analysis),** cognitive computing (e.g., SIRI)

Example 1

Using Machine Learning to Predict Risk of Opioid Overdose in Medicare

JAMA Network | Open™

Original Investigation | Substance Use and Addiction

Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions

Wei-Hsuan Lo-Ciganic, PhD; James L. Huang, PhD; Hao H. Zhang, PhD; Jeremy C. Weiss, MD, PhD; Yonghui Wu, PhD; C. Kent Kwoh, MD; Julie M. Donohue, PhD; Gerald Cochran, PhD; Adam J. Gordon, MD, MPH; Daniel C. Malone, PhD; Courtney C. Kuza, PhD; Walid F. Gellad, MD, MPH

More details, see JAMA Network Open. 2019;2(3):e190968

Rationale, Scientific Question & Methods

Rationale

- Current approaches to identifying individuals at high risk for opioid overdose target many patients who are not truly at high risk

Question

- Can machine-learning approaches more accurately predict opioid overdose risk among fee-for-service Medicare beneficiaries?

Methods

- A prognostic study using a 5% sample of claims data

560,057 fee-for-service beneficiaries without cancer who had ≥1 opioid fills

Randomly splitting into training, testing and validation samples

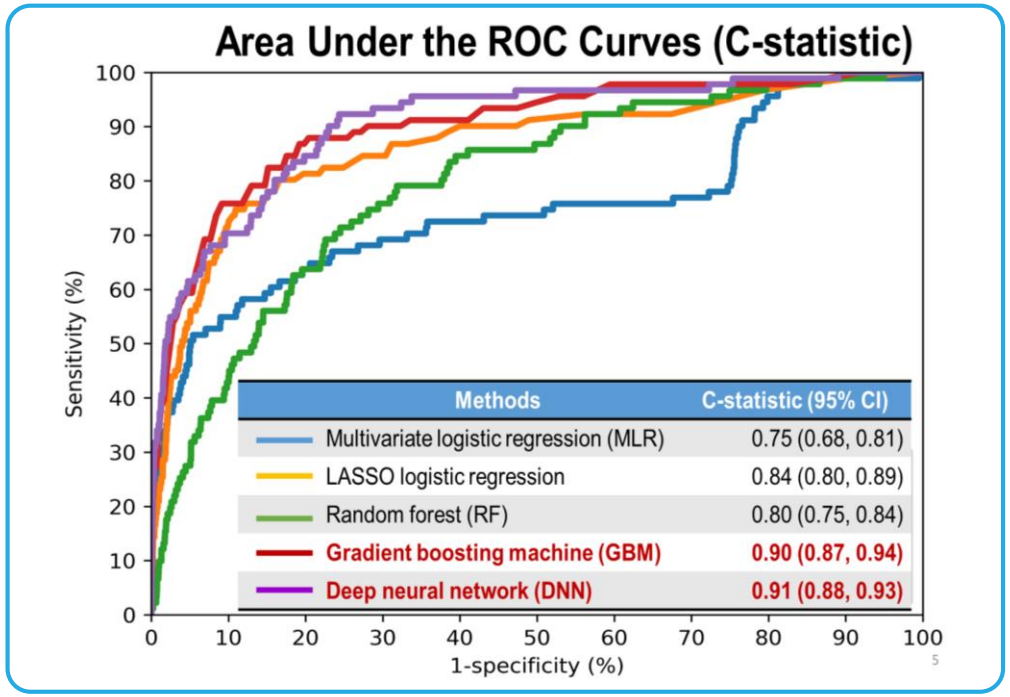
268 potential predictors (e.g., health status) measured in 3-month windows

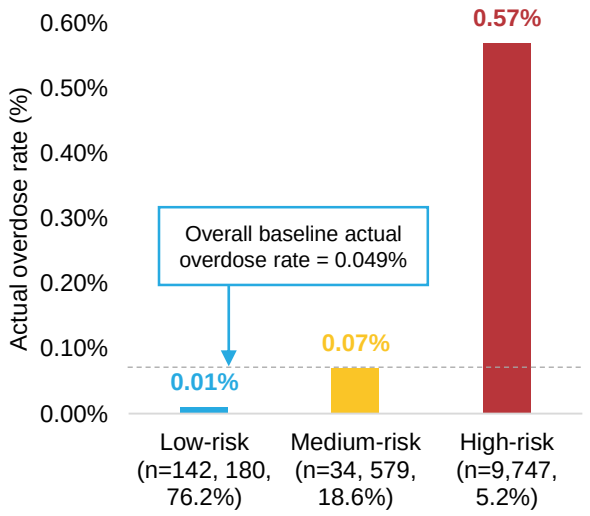
Primary outcome: any opioid overdose diagnosis in the subsequent 3 months

5 machine-learning approaches (e.g., deep neural network)

Prediction performance evaluation (e.g., C-statistics, number needed to evaluate)

Key Results & Main Conclusions





High-risk:
Overdose rate: 10 times higher than the overall rate Capture 62% patients with actual overdose; NNE=174

Medium-risk:
Overdose rate: 1.42 times higher than the overall rate Capture 29% patients with actual overdose; NNE=1,337

Low-risk:
actual overdose rate < overall rate (Almost everyone had non-overdose)

- **Low-risk** (predicted score below optimized score threshold)
- **Medium-risk** (between the optimized score threshold and the top 10th percentile score)
- **High-risk** (predicted score in the top 5th percentile)

≥90% of overdose episodes occurred in the high- and medium-risk subgroups

- Low positive predictive values, given the rare overdose outcome

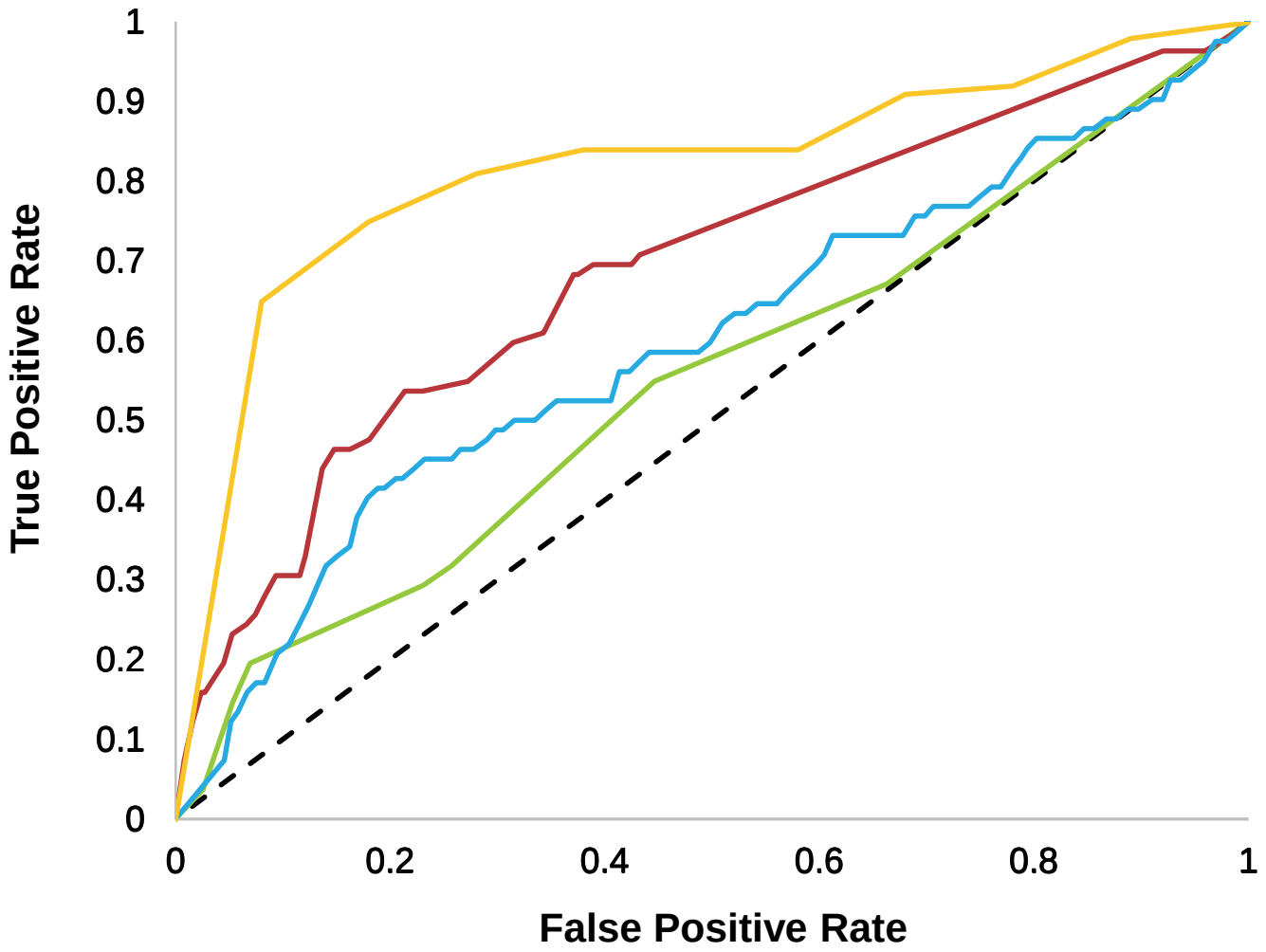
Machine learning algorithms appear to perform well for risk prediction and stratification of opioid overdose

More details, see JAMA Network Open. 2019;2(3):e190968

Example 2

Prediction of 90-day Hospitalization in a Health Plan

Comparison of ROC Curves – Hospitalization w/in 90 days



TreeNet – HRA Questions

0.83

Stepwise Logistic – HRA questions

0.74

Logistic w/all HRA questions

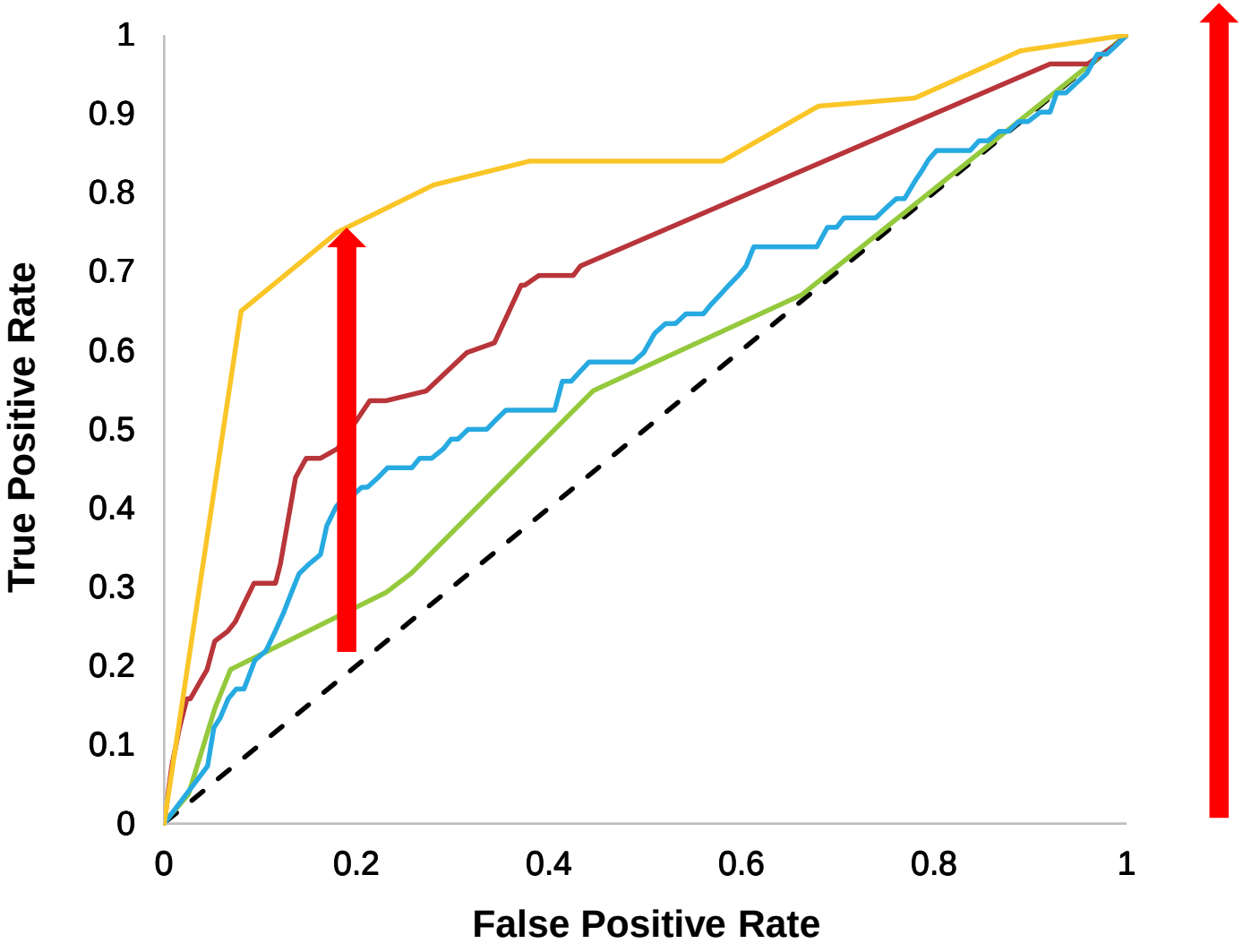
0.65

Demographics + Self-reported Health

0.56

A difference of 0.025 in c-stat is considered clinically significant. See Br J Anaesth. 2001 Jun;86(6):822-7.

Economic gains for predictive analytics



TreeNet – HRA Questions

$1.8M of prevented hospitalization expense relative to baseline

Stepwise Logistic – HRA questions

$473K of prevented hospitalization expense relative to baseline

Logistic w/all HRA questions

$247K of prevented hospitalization expense relative to baseline

Demographics + Self-reported Health

Baseline

Applying national health plan turnover rates to a population of 400K beneficiaries

When to Use Machine Learning for Prediction?

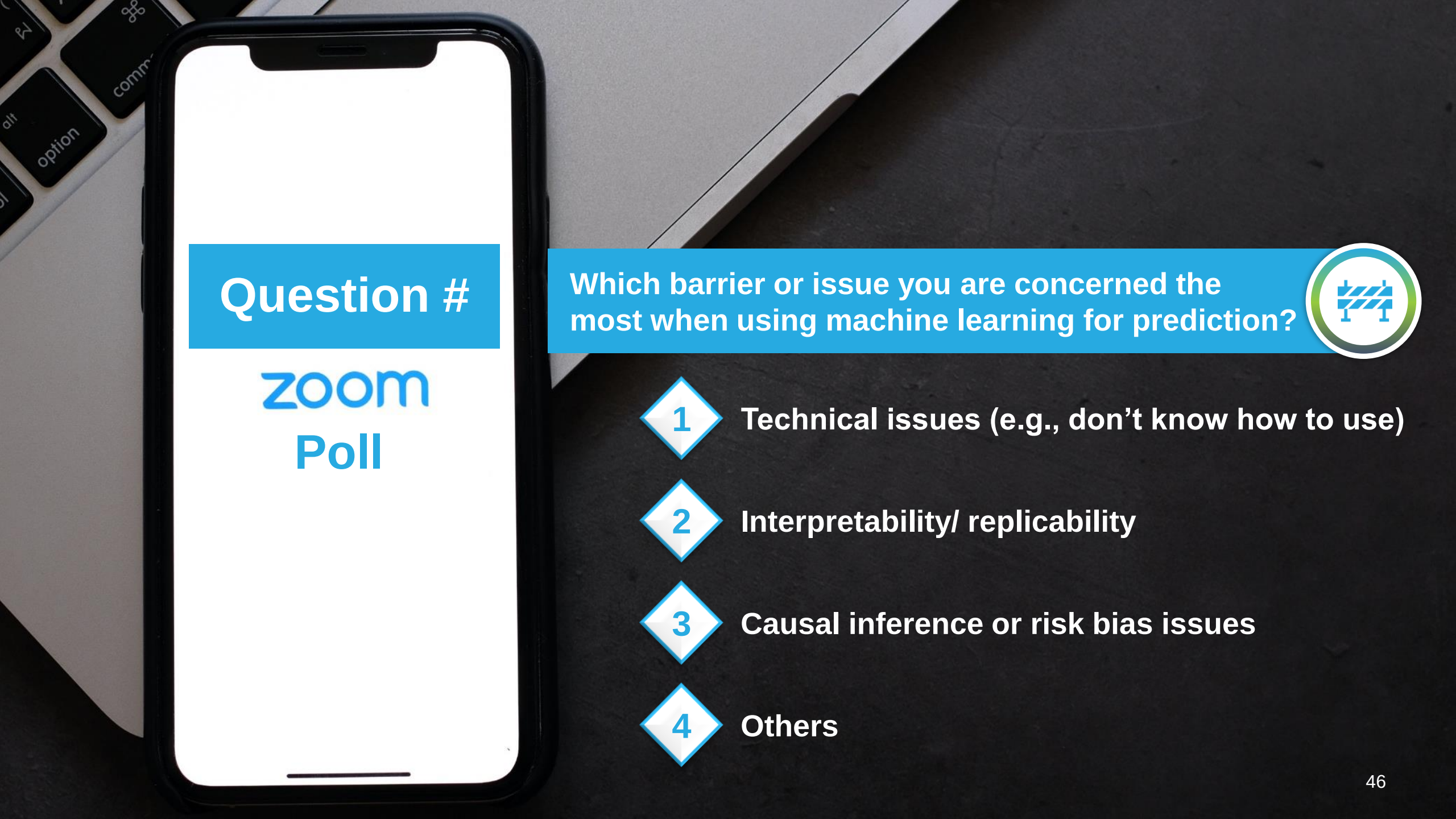
Problematic using traditional analytics, e.g., high-dimensional, highly correlated data

- Required restrict assumptions (e.g., linearity)
- Limitations of handling complex interactions, missing data, outliers
- Require to identify arbitrary thresholds or categories for continuous variables
- Complimentary to traditional analytics, not replacement!!

Problems are not easily codified by simple logical rules

- Complex interaction, hidden patterns

Accuracy is more important than interpretability



Which barrier or issue you are concerned the most when using machine learning for prediction?

1. Technical issues (e.g., don't know how to use)
2. Interpretability/ replicability
3. Causal inference or risk bias issues
4. Others

Q&A session

Danny Yeh, PhD,
Genentech, Head of Health Policy and Systems Research, US Medical Affairs, yehw@gene.com

Agustin Marquez, MS,
nference, Vice President of Product, BD and Marketing, agustin@nference.net

Danny Sheinson, PhD,
Genentech, Principal Data Scientist, US Medical Affairs, sheinsod@gene.com

Jenny Lo Ciganic PhD,
University of Florida, Assistant Professor, Pharmaceutical Outcomes and Policy, wlociganic@cop.ufl.edu

Backup

Issues to ponder when using machine learning for prediction

How to measure prediction performance? Which metrics you should use to inform clinical utility?

- Under what circumstance that C-statistic is not informative?

Causal inference vs. prediction are not the same

Implementation issues: any risk bias or ethical concerns when applying risk scores?