William Crown, PhD
Brandeis University
Waltham, MA, USA

Virtual

ISPOR Europe 2021 Spotlight 1

In "Deep" Thought: Advancing Machine Learning Methods in HEOR

ISPOR Europe 2021 Spotlight 1

In "Deep" Thought: Advancing Machine Learning Methods in HEOR

William Crown, PhD
Brandeis University
Waltham, MA, USA

Vivek Rudrapatna, MD, PhD
University of California
School of Medicine
San Francisco, CA, USA

Jenna Reps, PhD
Janssen Pharmaceuticals, Inc.
Raritan, NJ, USA

Karin Groothuis-Oudshoorn, PhD
University of Twente
Enschede, Netherlands

Agenda

16:00	Welcome and Background	Bill Crown, PhD
16:04	Machine Learning POV; Natural Language Processing; IPD medical data	Vivek Rudrapatna, MD, PhD
16:34	Causal Inference POV; Examples from Odyssey Collaboration	Jenna Reps, PhD
17:04	Discussant	Karin Groothuis-Oudshoorn, PhD
17:15	Q & A	Bill Crown, PhD

Antitrust Compliance Statement

- ISPOR has a policy of strict compliance with both United States, and other applicable international antitrust laws and regulations.
- Antitrust laws prohibit competitors from engaging in actions that could result in an unreasonable restraint of trade.
- ISPOR members (and others attending ISPOR meetings and/or events) must avoid discussing certain topics when they are together including, prices, fees, rates, profit margins, or other terms or conditions of sale.
- Members (and others attending ISPOR meetings and/or events) have an obligation to terminate any discussion, seek legal counsel's advice, or, if necessary, terminate any meeting if the discussion might be construed to raise antitrust risks.
- The Antitrust policy is available on the ISPOR website.

Please mute your line.

Vivek Rudrapatna, MD, PhD
University of California
School of Medicine
San Francisco, CA, USA

Virtual

ISPOR Europe 2021 Spotlight 1

In "Deep" Thought: Advancing Machine Learning Methods in HEOR

UCSF

Machine Learning for Clinical Text Classification: A Case Study

Vivek Rudrapatna, MD, PhD
Assistant Professor, Director of Real-World Evidence Projects
Division of Gastroenterology; Bakar Computational Health Sciences Institute
University of California, San Francisco

Disclosures

- None relevant to this presentation

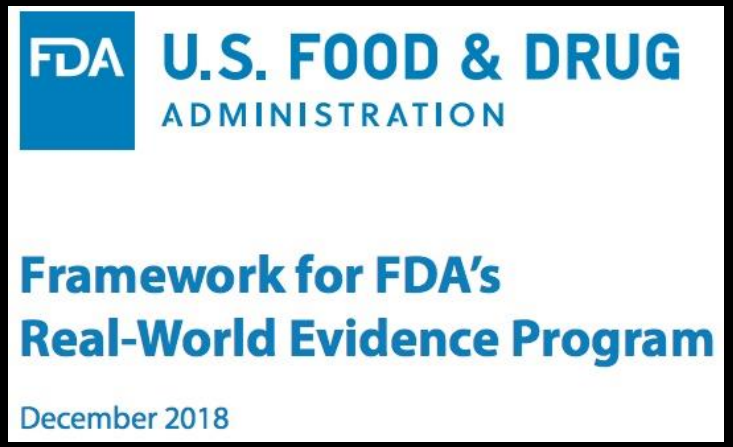
Learning objectives

- Understand the rationale for using machine learning (ML) for natural language processing tasks in the context of HEOR

- Understand the strengths and weaknesses of ML compared to biostatistical models

Real-world data is becoming increasingly important for every aspect of healthcare

- FDA and EMA are now using RWD to guide decision making on drug effectiveness and safety
  - Alecensa for lung cancer
  - Avelumab for Merkel cell cancer
  - Palbociclib for breast cancer in men
  - …



- Most examples have been in cancer and rare diseases, using hard endpoints
- Can this approach work for a much broader set of conditions, therapies, and outcome measures?

An HEOR case study in ulcerative colitis

- Pilot study:
  - Can we use EHR data to estimate Tofacitinib's effectiveness for UC?
  - Do our estimates match the results of pivotal/registrational trials?

Utility of routinely collected electronic health records data to support effectiveness evaluations in inflammatory bowel disease: a pilot study of tofacitinib

Vivek Ashok Rudrapatna [1,2] Benjamin Scott Glicksberg [1,3]
Atul Janardhan Butte [1,4]

- Challenge:
  - Disease activity measures and endpoints use patient- and physician-reported outcomes
  - They are typically captured in clinical notes, <u>not</u> as structured (analysis-ready) data

Table A. Mayo Score

Stool frequency*
- 0 = Normal no. of stools for this patient
- 1 = 1–2 stools more than normal
- 2 = 3–4 stools more than normal
- 3 = 5 or more stools more than normal

Rectal bleeding†
- 0 = No blood seen
- 1 = Streaks of blood with stool less than half the time
- 2 = Obvious blood with stool most of the time
- 3 = Blood alone passed

Findings of flexible proctosigmoidoscopy
- 0 = Normal or inactive disease
- 1 = Mild disease (erythema, decreased vascular pattern, mild friability)
- 2 = Moderate disease (marked erythema, absent vascular pattern, friability, erosions)
- 3 = Severe disease (spontaneous bleeding, ulceration)

Rudrapatna, Glicksberg, and Butte. *BMJ Health & Care Informatics*, 2021; www.fda.gov/media/99526/download

This problem is not unique to ulcerative colitis!

- Applies to any disease where improvement is measured by patients themselves…
  - PROs in depression, arthritis, insomnia, irritable bowel syndrome, etc
- …or by their treating physicians
  - progression-free survival, endoscopic severity, etc

- The common approach is to use proxies (hospitalizations, surgeries, medications) derived from existing structured datasets (claims, EHR)
  - But hard to validate, prone to confounding, measurement bias, etc

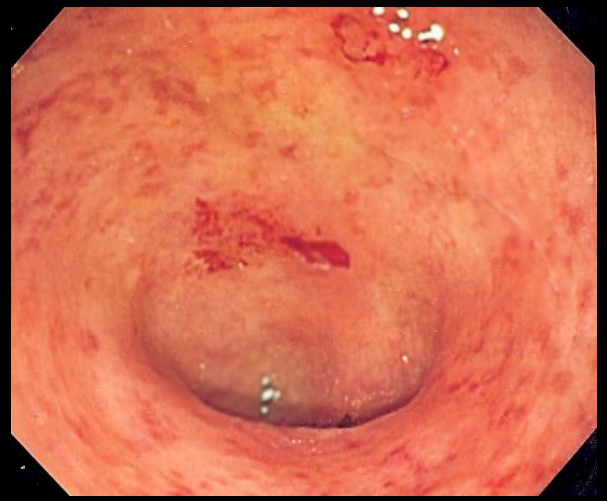
- To control confounding, characterize cohorts, and accurately measure exposures and outcomes, we need new tools for extracting information from clinical text!

S0818 Accurate Machine Classification of Ulcerative Colitis Mayo Subscores From Electronic Health Record Procedure Reports

Rudrapatna, Vivek MD, PhD[1]; Gupta, Saransh BS[2]; Mardirossian, Taline BS[2]; Narain, Rohan BS[2]; Mosenia, Arman MSE[1]; Butte, Atul MD, PhD[1] **Author Information**

The American Journal of Gastroenterology: October 2020 - Volume 115 - Issue - p S420
doi: 10.14309/01.ajg.0000705320.33817.a5

- Need to go from this…



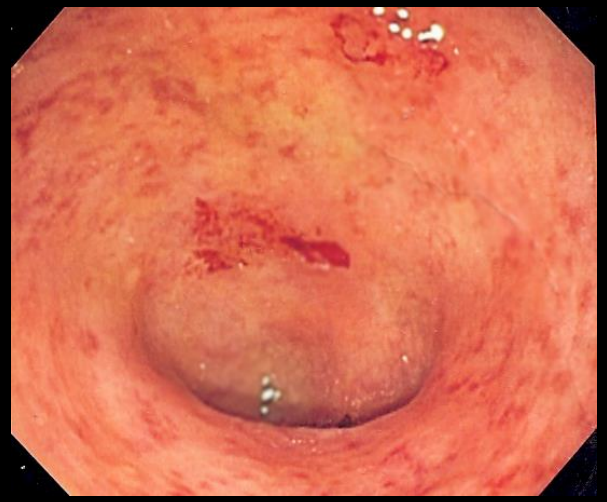
COLON FINDINGS: A diffuse circumferential patch of inflammation was found in the rectum, sigmoid colon, and descending colon. The mucosa was congested, edematous, erythematous, friable and had erosions. Moderate to severe colitis to splenic flexure. At this point it appears to stop. However, exam stopped because of severity of disease distal.

S0818 Accurate Machine Classification of Ulcerative Colitis Mayo Subscores From Electronic Health Record Procedure Reports

Rudrapatna, Vivek MD, PhD[1]; Gupta, Saransh BS[2]; Mardirossian, Taline BS[2]; Narain, Rohan BS[2]; Mosenia, Arman MSE[1]; Butte, Atul MD, PhD[1] **Author Information**

The American Journal of Gastroenterology: October 2020 - Volume 115 - Issue - p S420
doi: 10.14309/01.ajg.0000705320.33817.a5

- Need to go from this…to this!

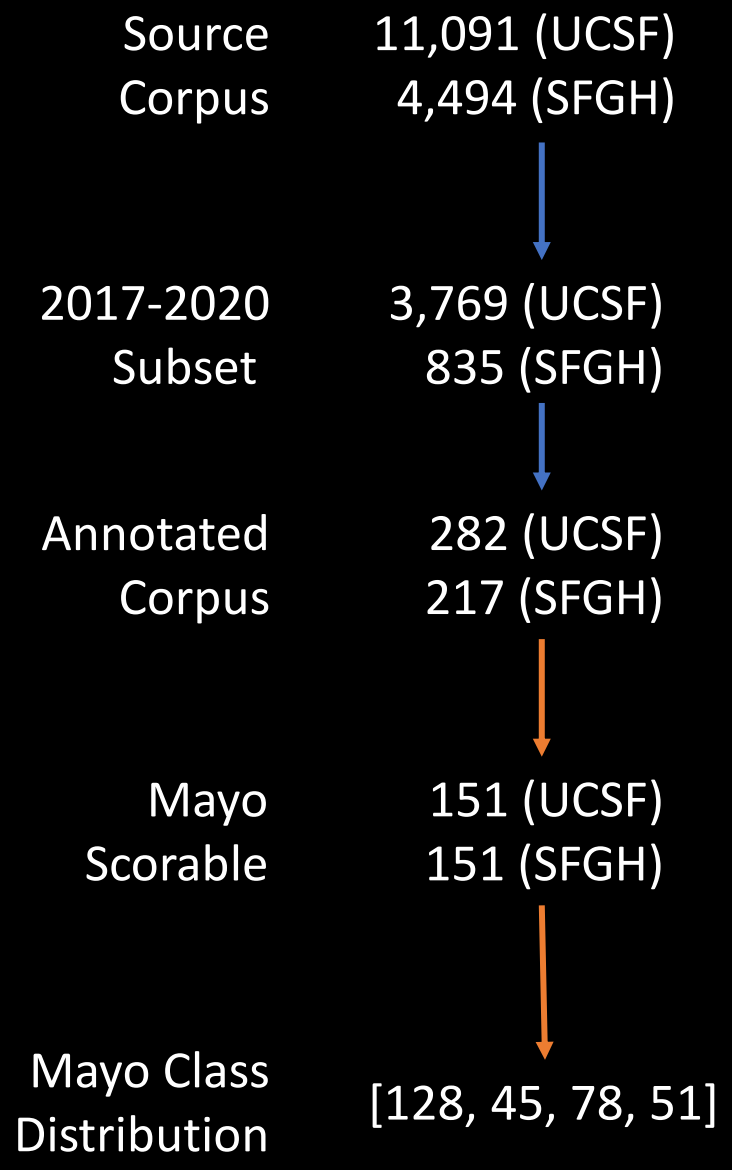


Findings on Endoscopy:
0 = Normal or inactive disease.
1 = Mild disease (erythema, decreased vascular pattern).
2 = Moderate disease (marked erythema, lack of vascular pattern, any friability, erosions).
3 = Severe disease (spontaneous bleeding, ulceration).

Mayo 3

Methods

- Extracted endoscopy reports from two centers
  - Different patients (tertiary care vs safety-net), providers, equipment, reporting software
- Defined an annotation protocol, confirmed IRR
- Sequentially annotated a uniform sample of notes to obtain 150 scorable reports/center
  - Scorable ~ not Crohn's disease, ileal pouch, etc.



Orange arrows denote ML prediction tasks

Methods

- The task was framed as one of text classification
  - Whether or not the report was Mayo scorable (binary)
  - What the Mayo score was (multiclass)

- Predictive features:
  - Expert-driven features (“erosion*”, “ulcer*”, “edema*”)
  - *Bag of Words*: Every uni/bi/trigram in the source notes (N ~ 50K)
    - E.g. “mucosa”, “was congested”, “friable and ulcerated”, …

- Trained a series of predictive models based on these features
  - Baseline models trained in *Python* using *sklearn,* default hyperparameters
  - Model selection using 5-fold cross-validation
  - Evaluated once using a hold-out test set (20% of the data), stratified on the outcome

Results: Binary Classification (baseline)

Classifier	Accuracy (%)	F1 Score(%)
Naive Bayes	74	73
KNN	69	66
Logistic Regression	89	89
Decision Tree	92	93
Random Forest	89	89
ExtraTrees	89	89
AdaBoost	96	96
SVM	60	45
XGBoost	60	45
LightGBM	97	98

Results: Multiclass Prediction (baseline)

Classifier	Accuracy (%)	F1 Score(%)
Naive Bayes	64	63
KNN	41	34
Logistic Regression	72	72
Decision Tree	69	68
Random Forest	64	58
ExtraTrees	70	69
AdaBoost	66	65
SVM	31	15
XGBoost	31	21
LightGBM	77	77

Improving learning efficiency with AutoML

Best Classifier	Accuracy (%)	F1 Score(%)
LightGBM (Binary)	97	98
AutoML RF (Binary)	97	98
LightGBM (Multi)	68	75
AutoML XGBoost (Multi)	97	97

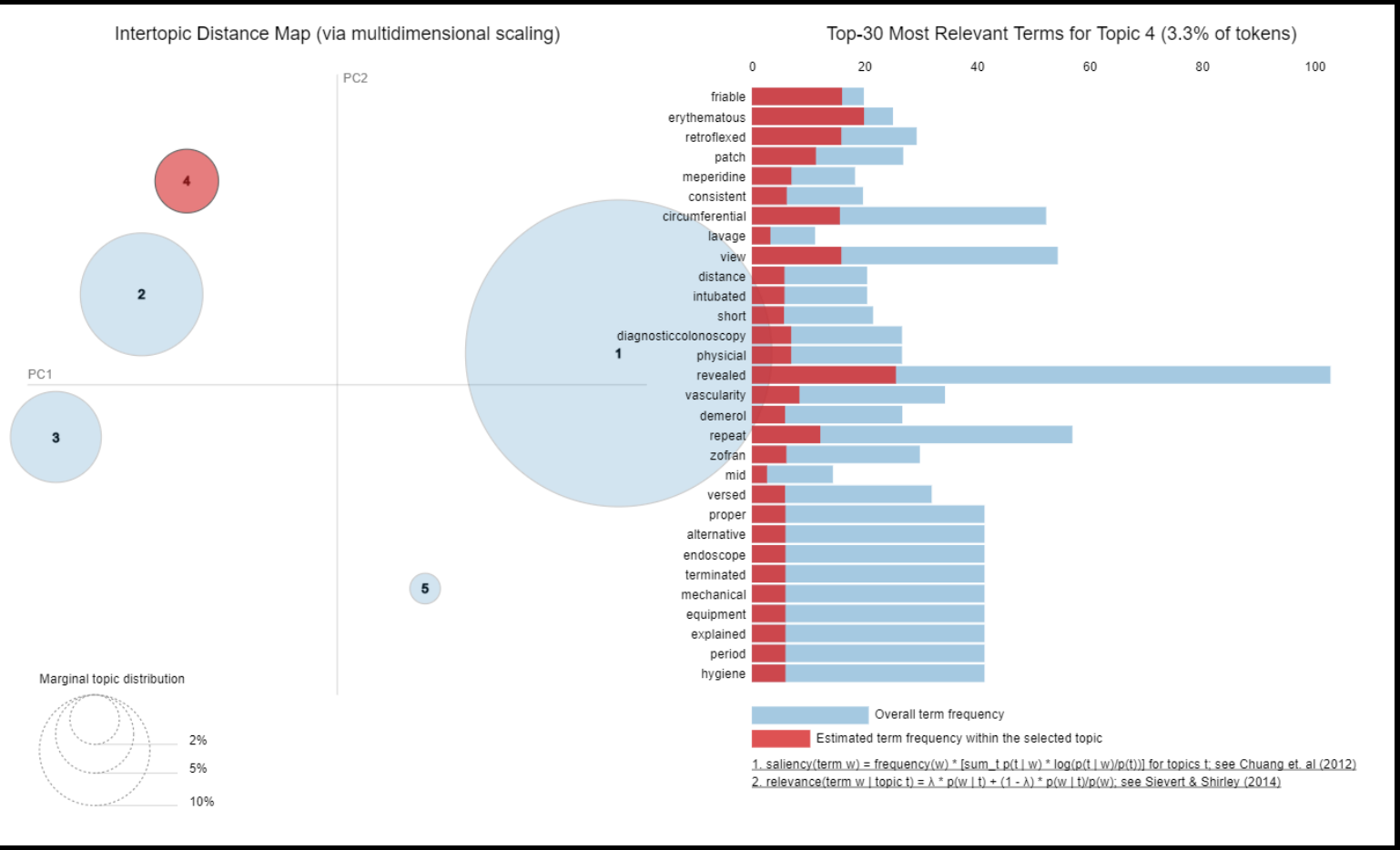
- Many ML models are very sensitive to features and hyperparameters
  - Requires lots of upfront engineering, tuning to get the best performance
- AutoML = software for automating the ML modeling process itself!
  - Hyperparameter optimization
  - Ensembling (~ bagging/boosting)
  - Stacking (~ deep learning)

Erickson, Mueller, Shirkov, Zhang, Larroy, Li, and Smola. *arXiv* 2020; auto.gluon.ai

Model Interpretability

- Did it learn what we expected it to learn?
  - This is a case where we happen to know the data generating function
    - Bleeding and ulcerated mucosa -> Mayo 3
    - Erosions, Friability -> Mayo 2
    - …
  - Are these black-box models useful for performing feature selection, discovery?

Model Interpretability



- Topic modeling using Latent Dirichlet Allocation
  - Generative latent variable model
  - Model is fit according to the assumption that there are 5 classes that give rise to the data
  - Assigns tokens to each class

What does it take to get models like this?

- Computation: ~2 hours on a standard laptop. No GPUs required!

- Data: not as much as you would expect (esp if it were a deep learning model)

Binary Classifier	# of Reports	F1 Score
AutoML LightGBM	400 (100%)	96%
AutoML XGBoost	320 (80%)	97%
AutoML LightGBM	240 (60%)	84%

Multi Classifier	# of Reports	F1 Score
AutoML XGBoost	242 (100%)	97%
AutoML CatBoost	194 (80%)	67%
AutoML XGBoost	145 (60%)	57%

- Data diversity: When training on one health system's data and evaluating generalizability to the second, we lose ~10% on accuracy/F1

Discussion

- AutoML may be a practical and efficient solution for many real-world ML tasks, including for HEOR and particularly in resource-limited settings.

- Deep learning is not required for every NLP task[1]
  - For 'lower level' NLP tasks (keywords ± negation), classical methods may be sufficient

- ML vs. biostatistical models
  - ML when you're dealing with many 'nuisance' parameters, prediction >> interpretability
    - Does better with raw/sensory data for which feature engineering is needed, can be automated
  - Biostats superior for causal inference, requires the analyst to specify the causal model (DAG)
  - ML can complement biostat models as an informal test of possible model misspecification
  - Black-box nature of ML -> prone to algorithmic bias
    - ML Interpretability remains a major open challenge
    - ML feature importance ≠ causal features!

1. Tabular data: deep learning is not all you need. https://arxiv.org/abs/2106.03253

Acknowledgements

- Anna Silverman*, Balu Bhasuran*, Arman Mosenia, Rohan Narain, Taline Mardirossian, Saransh Gupta, Atul Butte

- The Atul Butte Laboratory
- Bakar Computational Health Sciences Institute
- UCSF Center for Colitis and Crohn's Disease, Division of Gastroenterology
- Funding: National Institutes of Health, Janssen Inc, Alnylam Inc

*: Denotes equal contribution

We are building a Center for Real-World Evidence at the University of California Health System!

- 6 medical centers, 7M+ patients
- OMOP-formatted EHR data + machine redacted clinical notes
  - 100M+ at UCSF since 2012 (6B clinical concepts!)
  - Customized BERT model for natural language inference
- We routinely collaborate with the FDA and biopharmaceuticals on RWE projects
- Reach out if you want to discuss a collaboration! evidence@ucsf.edu

UNIVERSITY OF CALIFORNIA

UCSF

UCSF

Thank you!

@vivicality

Jenna Reps, PhD
Janssen Pharmaceuticals, Inc.
Raritan, NJ, USA

Virtual

ISPOR Europe 2021 Spotlight 1

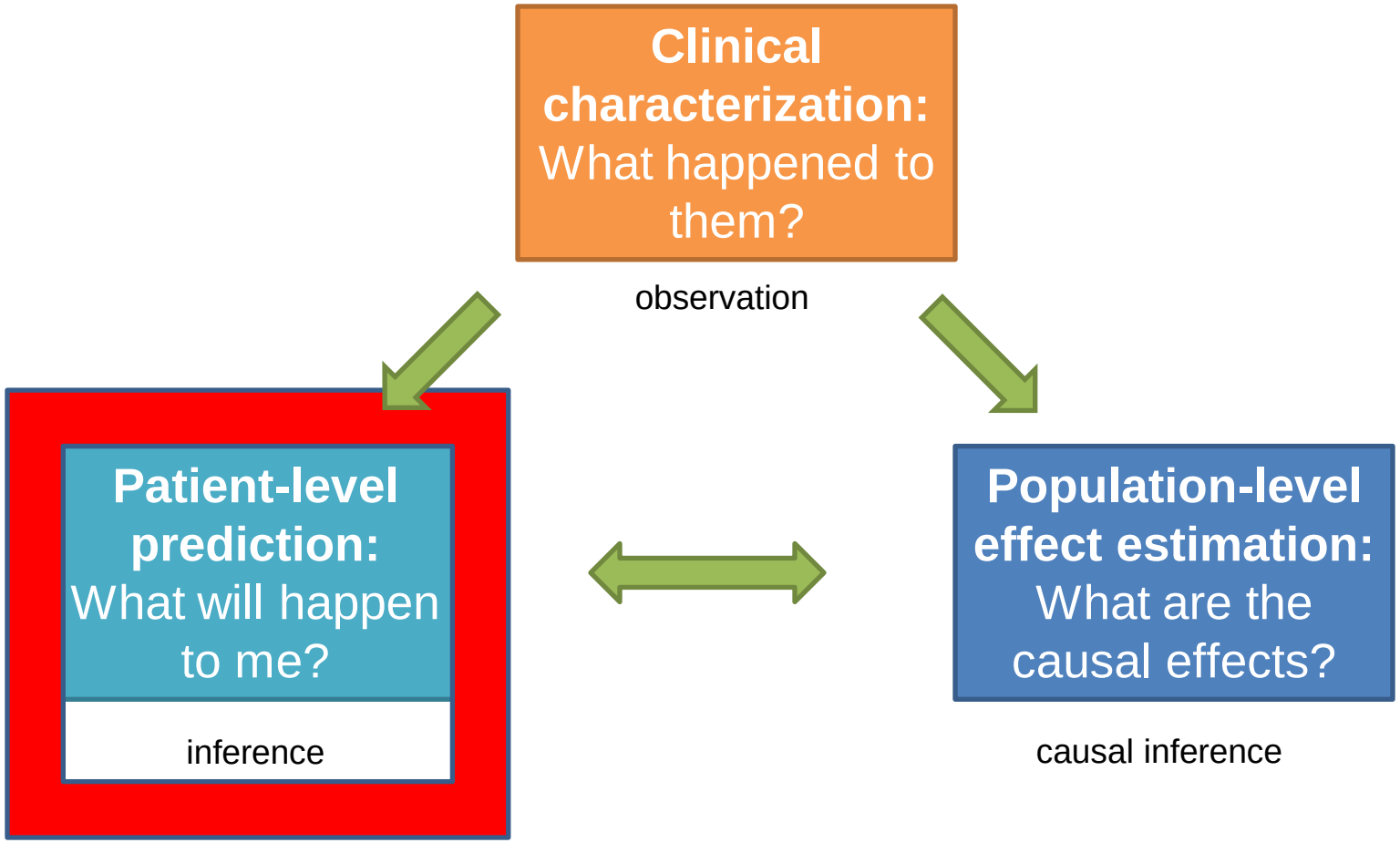
In "Deep" Thought: Advancing Machine Learning Methods in HEOR

Best Practices for Patient-level Prediction using observational data

ISPOR Europe 2021 – 30th November

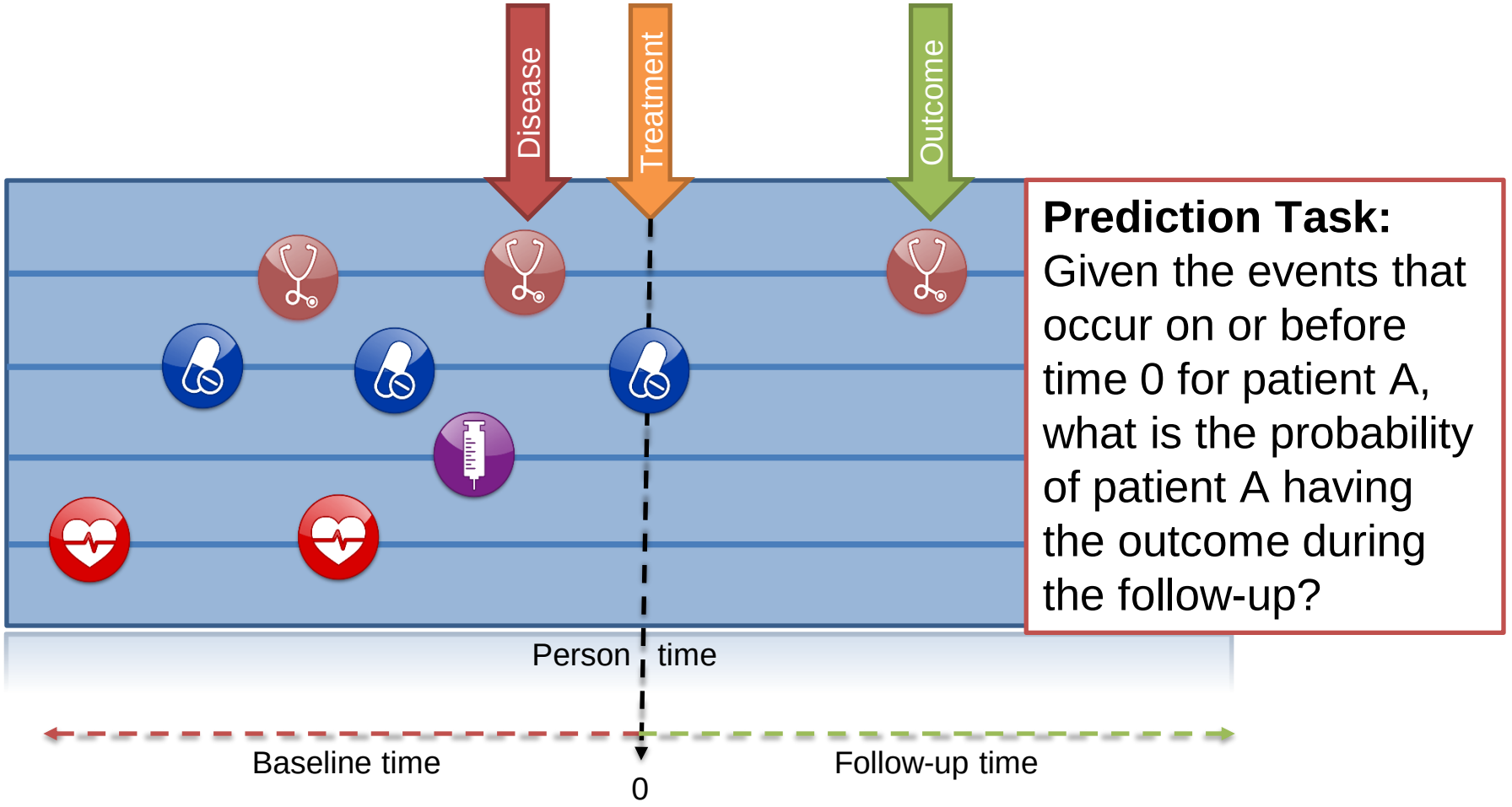
Jenna Reps, Janssen R&D / OHDSI

What prediction does and what it does not do



We can use observational data to develop prediction models

Person A



Prediction Task:
Given the events that occur on or before time 0 for patient A, what is the probability of patient A having the outcome during the follow-up?

What questions can prediction answer?

Type	Structure	Example
Disease onset and progression	Amongst patients who are newly diagnosed with **<insert your favorite disease>**, which patients will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?	Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?
Treatment choice	Amongst patients with **<indicated disease> who are treated with either <treatment 1> or <treatment 2>**, which patients were treated with **<treatment 1>** (**on day 0)**?	Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)
Treatment response	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will **<insert desired effect>** in **<time window>**?	Which patients with T2DM who start on metformin stay on metformin after 3 years?
Treatment safety	Amongst patients who are new users of **<insert your favorite drug>**, which patients will experience **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?	Among new users of warfarin, which patients will have GI bleed in 1 year?
Treatment adherence	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will achieve **<adherence metric threshold>** at **<time horizon>**?	Which patients with T2DM who start on metformin achieve >=80% proportion of days covered at 1 year?

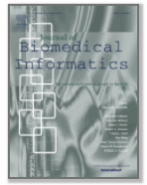
Prediction does get used in causal inference

Type	Structure	Example
Disease onset and progression	Amongst patients who are newly diagnosed with **<insert your favorite disease>**, which patients will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?	Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?
Treatment ch oice	Amongst patients with **<indicated disease> who are treated with either <treatment 1> or <treatment 2>**, which patients were treated with **<treatment 1>** (**on day 0**)?	Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)
Treatment response	Amongst patients who are new users of **<favorite chronically-used drug>**, which will **<insert desired effect>** in **<time win**	
Treatment safety	Amongst patients who are new users of **<favorite drug>**, which patients will experi **favorite known adverse event from the profile>** within **<time horizon following**	
Treatment adherence	Amongst patients who are new users of **<favorite chronically-used drug>**, which achieve **<adherence metric threshold> horizon>**?	

ELSEVIER

Journal of Biomedical Informatics

Volume 56, August 2015, Pages 356-368



A supervised adverse drug reaction signalling framework imitating Bradford Hill's causality considerations

Jenna Marie Reps [a], Jonathan M. Garibaldi [a], Uwe Aickelin [a], Jack E. Gibson [b], Richard B. Hubbard [b]

Unfortunately, most models fail to make an impact

Making Machine Learning Models Clinically Useful

Recent advances in supervised machine learning have improved diagnostic accuracy and prediction of treatment outcomes, in some cases surpassing the performance of clinicians.[1] In supervised machine learning, a mathematical function is constructed via automated analysis of training data, which consists of input features (such as retinal images) and output labels (such as the grade of macular edema). With large training data sets and minimal human guidance, a computer learns to generalize from the information contained in the training data. The result is a mathematical function, a model, that can be used to map a new record to the corresponding diagnosis, such as an image to grade macular edema. Although machine learning–based models for classification or for predicting a future health state are being developed for diverse clinical applications, evidence is lacking that deployment of these models has improved care and patient outcomes.[2]

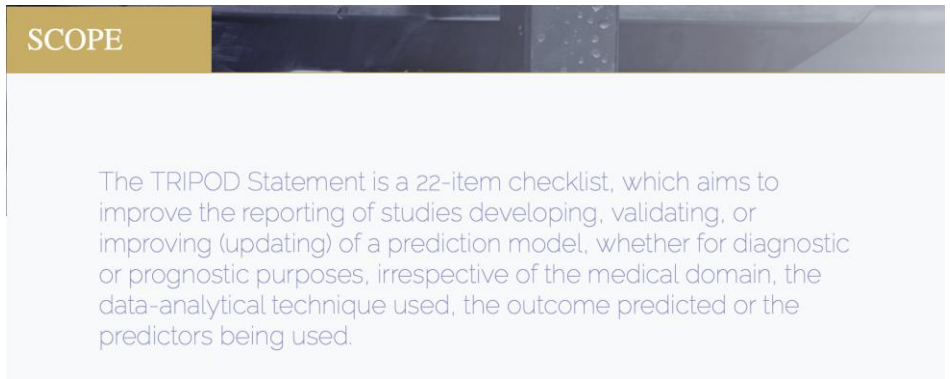
There have been a lot of reporting/development guidelines

Some of the guidelines:
PROGRESS – recommendations/considerations for model development
TRIPOD – how to transparently report prediction models
PROBAST – how to assess bias when reviewing prediction models

Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research

Ewout W. Steyerberg, Karel G. M. Moons, Danielle A. van der Windt, Jill A. Hayden, Pablo Perel, Sara Schroter, Richard D. Riley, Harry Hemingway, Douglas G. Altman, for the PROGRESS Group

Published: February 5, 2013 • https://doi.org/10.1371/journal.pmed.1001381

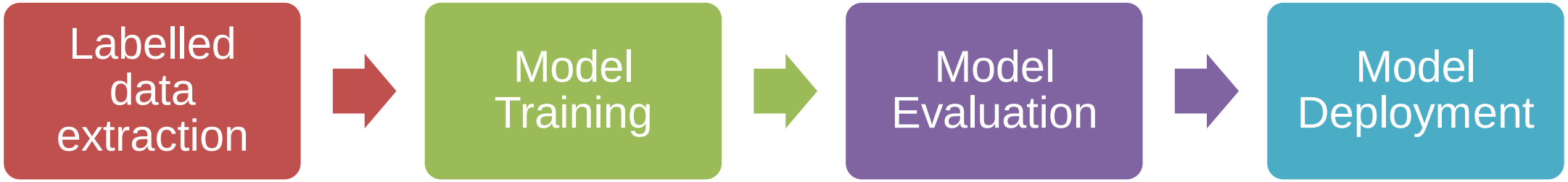


PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration FREE

Karel G.M. Moons, PhD*, Robert F. Wolff, MD*, Richard D. Riley, PhD, Penny F. Whiting, PhD, ... View all authors +

The identified issues are 1) lack of clinical usefulness and 2) poor design

Model Design Process:



- Numerous points along the process where a design decision must be made…
- **but** little research on the impact of each design choice

First: make sure any prediction you do has clinical value

A nice example:

A framework for making predictive models useful in practice

Kenneth Jung, Sehj Kashyap, Anand Avati, Stephanie Harman, Heather Shaw, Ron Li, Margaret Smith, Kenny Shum, Jacob Javitz, Yohan Vetteth ... Show more

Journal of the American Medical Informatics Association, Volume 28, Issue 6, June 2021, Pages 1149–1158, https://doi.org/10.1093/jamia/ocaa318

Published: 22 December 2020 **Article history ▾**

Second: make sure you use the correct design when developing models

- **How?** There is currently a lack of guidance

- **Solution:** We have a community of researchers that are developing best practices by doing empirical experiments

OHDSI is a worldwide collaboration

We use observational healthcare data

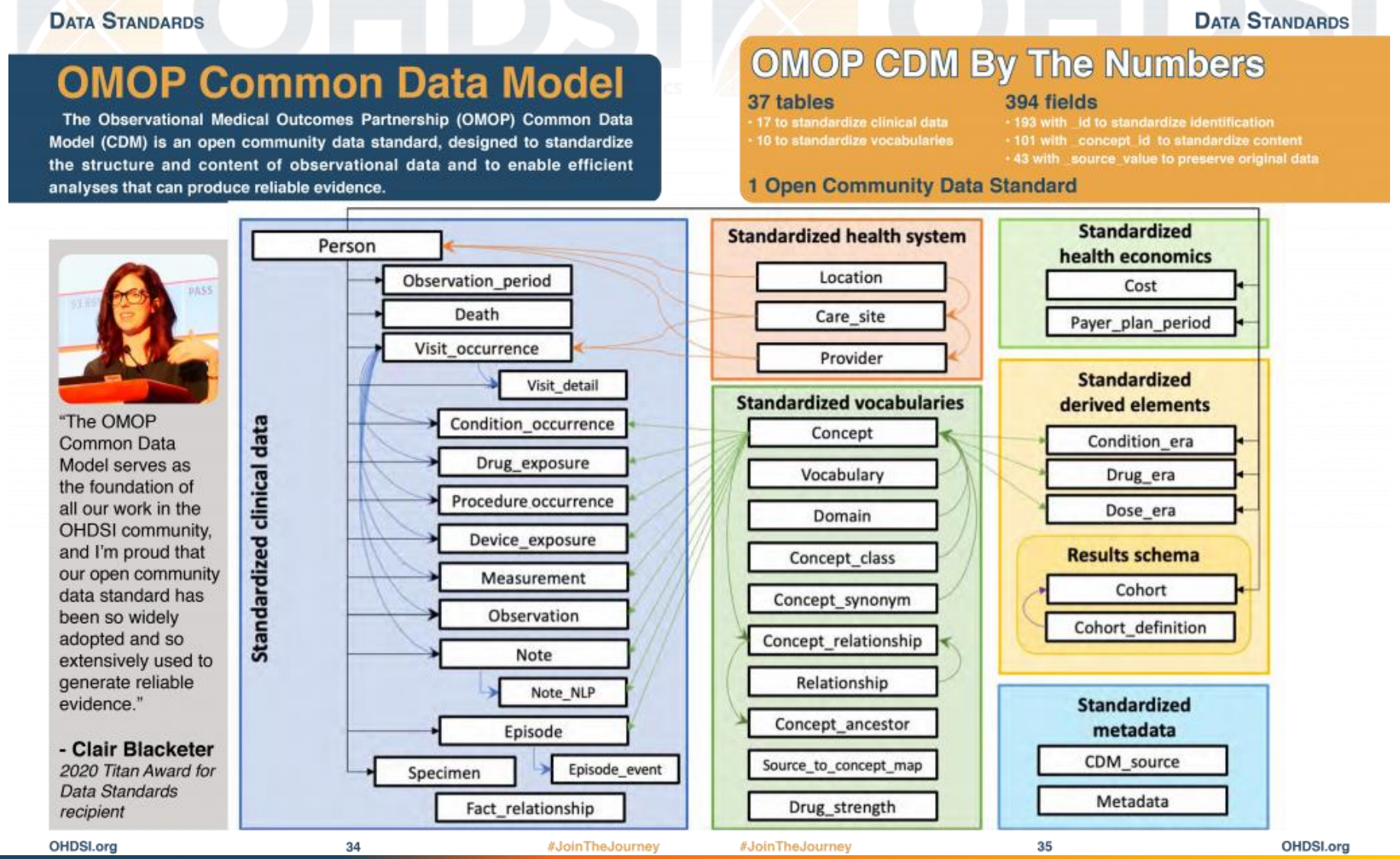
Our Mission

To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

https://ohdsi.org

A common data model

- OHDSI uses a standardized data structure known as the OMOP CDM
- This lets us develop tools/software that can be shared between researchers



Standardized patient-level prediction framework

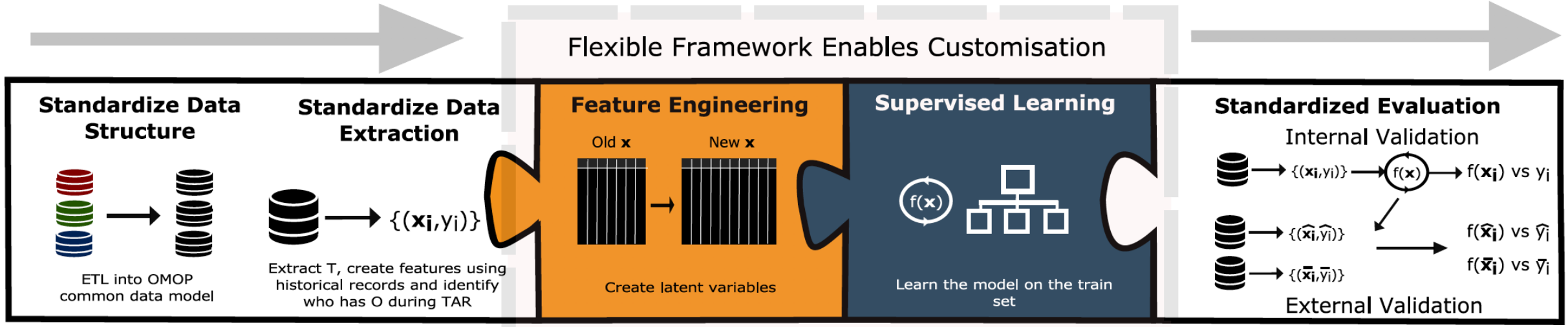
Amongst patients with **<T>**, which patients will go on to have **<O>** within **<time-at-risk>**?

Component	Description
Target cohort (T)	The people you want to do the prediction for (and the index date)
Outcome cohort (O)	The event being predicted
Time-at-risk (TAR)	The time period relative to the T index date to predict the occurrence of O within

Standardized patient-level prediction framework

Patient-Level Prediction Framework

We have developed an end-to-end framework for developing and validating clinically useful prediction models using observational data



Includes a library of methods that can be expanded
Methods: random forest, KNN, deep learning, gradient boosting machines, adaboost, decision tree and regularized logistic regression

Patient-level prediction R Package

Open Source: https://github.com/OHDSI/PatientLevelPrediction

Website: https://ohdsi.github.io/PatientLevelPrediction/

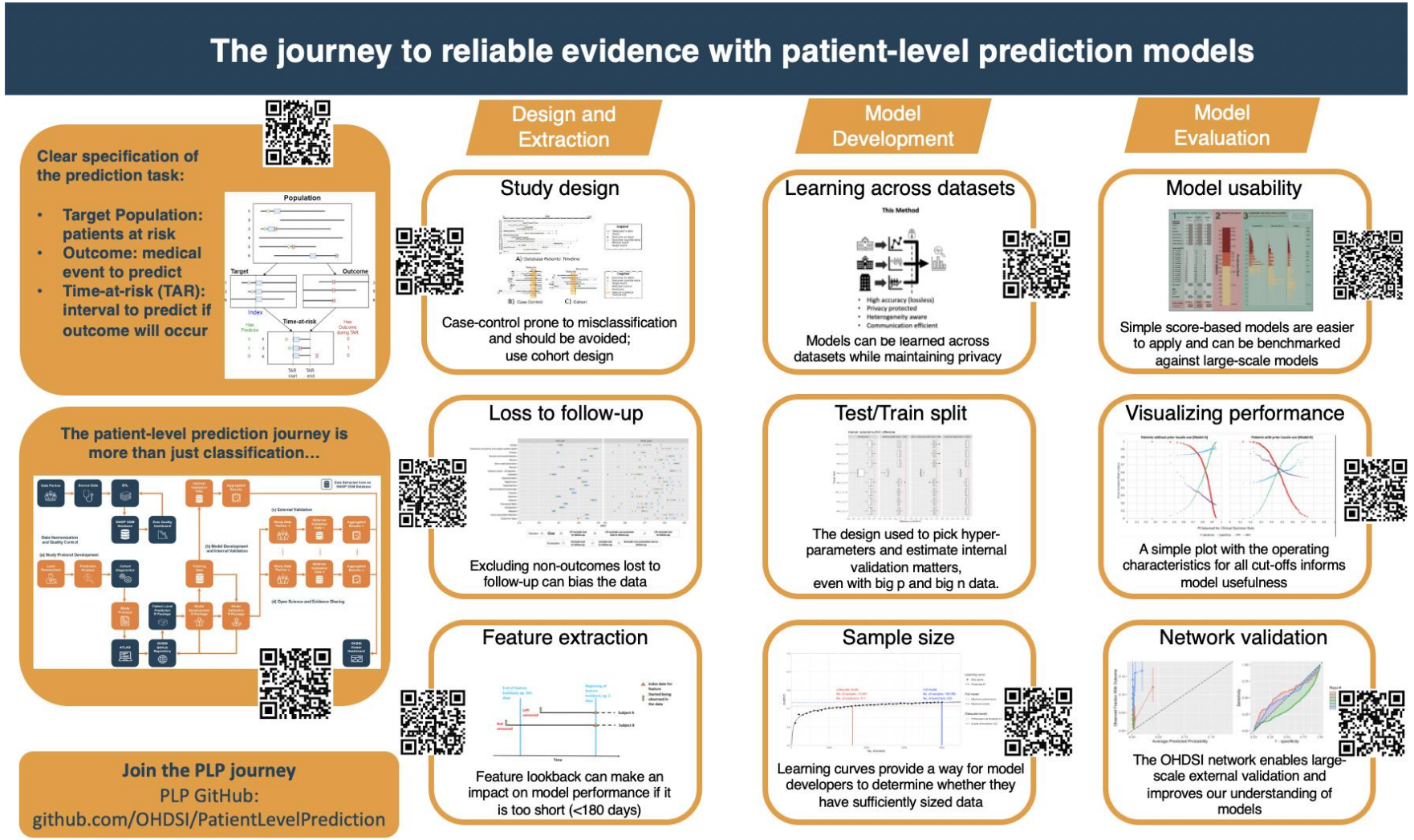
Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data

Jenna M Reps ✉, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, Peter R Rijnbeek

Journal of the American Medical Informatics Association, Volume 25, Issue 8, August 2018, Pages 969–975, https://doi.org/10.1093/jamia/ocy032

Published: 27 April 2018 **Article history** ▾

Using OHDSI standardizations we can investigate best practices



The process of extracting labelled data

Labelled data extraction

- How to create the features?
- How to define the outcome?
- How to handle missing features or missing outcome

Research | Open Access | Published: 16 August 2021

Design matters in patient-level prediction: evaluation of a cohort vs. case-control design when developing predictive models in observational healthcare datasets

Jenna M. Reps, Patrick B. Ryan, Peter R. Rijnbeek & Martijn J. Schuemie

Journal of Big Data **8**, Article number: 108 (2021) | Cite this article

An empirical analysis of dealing with patients who are lost to follow-up when developing prognostic models using a cohort design

Jenna M. Reps, Peter Rijnbeek, Alana Cuthbert, Patrick B. Ryan, Nicole Pratt & Martijn Schuemie

BMC Medical Informatics and Decision Making **21**, Article number: 43 (2021) | Cite this article

1326 Accesses | **3** Altmetric | Metrics

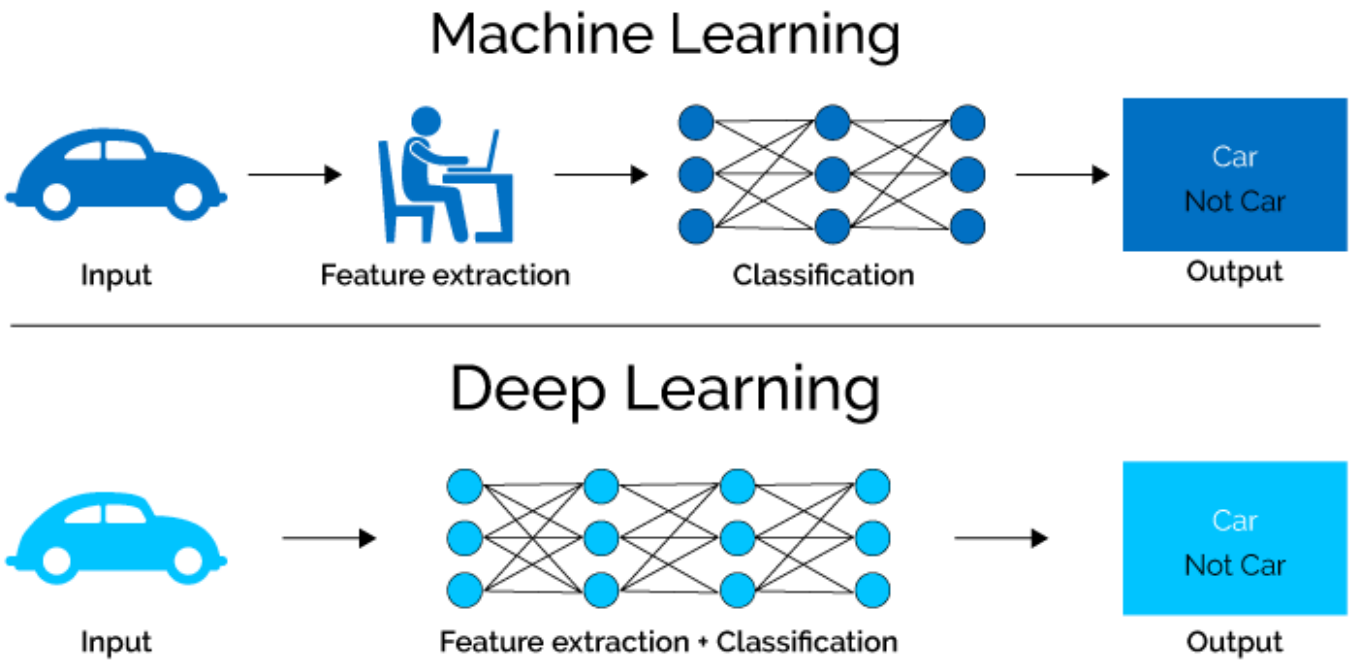
Research | Open Access | Published: 28 August 2021

Evaluating the impact of covariate lookback times on performance of patient-level prediction models

Jill Hardin & Jenna M. Reps

BMC Medical Research Methodology **21**, Article number: 180 (2021) | Cite this article

Machine learning such as deep learning could learn the features…



Research | Open Access | Published: 28 August 2021

Evaluating the impact of covariate lookback times on performance of patient-level prediction models

Jill Hardin & Jenna M. Reps

BMC Medical Research Methodology **21**, Article number: 180 (2021) | Cite this article

We can improve performance by learning the features using deep learning…

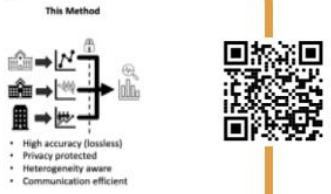
The process of learning the prediction model

Model Training

- What classifier?
- How to pick the hyper-parameters?
- Should we learn across datasets?
- How much data do we need?

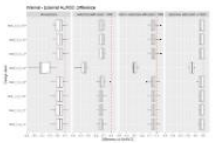
Model Development

Learning across datasets



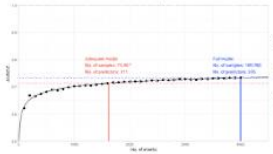
Models can be learned across datasets while maintaining privacy

Test/Train split



The design used to pick hyper-parameters and estimate internal validation matters, even with big p and big n data.

Sample size



Learning curves provide a way for model developers to determine whether they have sufficiently sized data

Lossless Distributed Linear Mixed Model with Application to Integration of Heterogeneous Healthcare Data

Chongliang Luo, Md. Nazmul Islam, Natalie E. Sheils, Jenna Reps, John Buresh, Rui Duan, Jiayi Tong, Mackenzie Edmondson, Martijn J. Schumie, Yong Chen

doi: https://doi.org/10.1101/2020.11.16.20230730

How little data do we need for patient-level prediction?

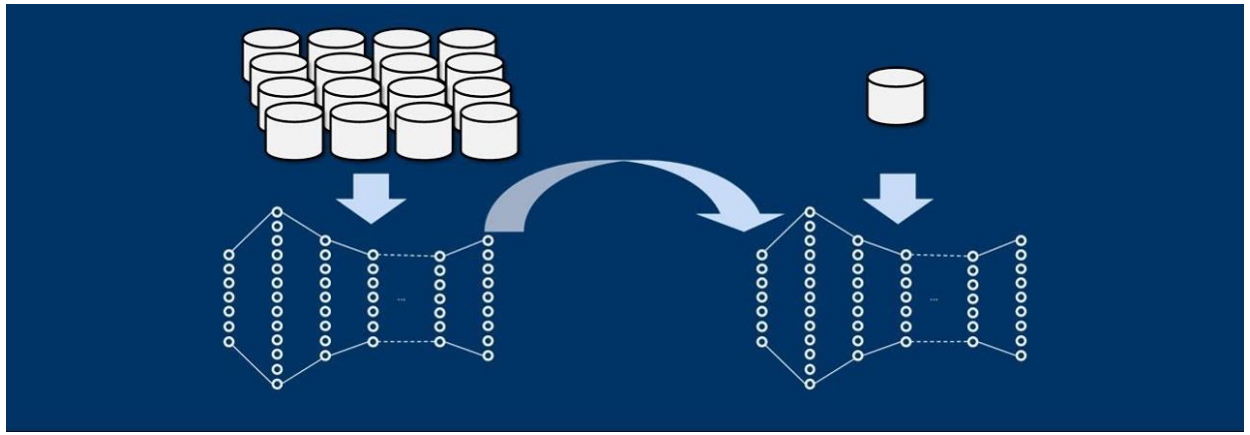
Luis H. John, Jan A. Kors, Jenna M. Reps, Patrick B. Ryan, Peter R. Rijnbeek

Objective: Provide guidance on sample size considerations for developing predictive models by empirically establishing the a
competing objectives of improving model performance and reducing model complexity as well as computational requirement
Materials and Methods: We empirically assess the effect of sample size on prediction performance and model complexity by g
prediction problems in three large observational health databases, requiring training of 17,248 prediction models. The adequ
sample size for which the performance of a model equalled the maximum model performance minus a small threshold value.
Results: The adequate sample size achieves a median reduction of the number of observations between 9.5% and 78.5% for t
0.02. The median reduction of the number of predictors in the models at the adequate sample size varied between 8.6% and
Discussion: Based on our results a conservative, yet significant, reduction in sample size and model complexity can be estima
if a researcher is willing to generate a learning curve a much larger reduction of the model complexity may be possible as su
variability.
Conclusion: Our results suggest that in most cases only a fraction of the available data was sufficient to produce a model clos
on the full data set, but with a substantially reduced model complexity.

Transfer learning is very powerful

- Why not learn latent features **across** datasets and then fine tune the model on the smaller local data?

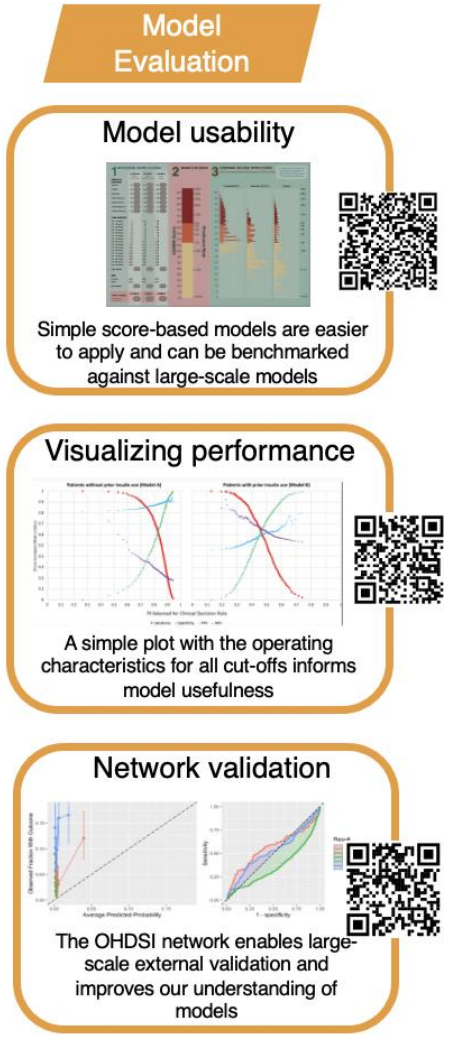
- This is where machine learning such as deep learning may really **improve** prediction



The process of evaluating the model

Model Evaluation

- How to get a fair performance estimate?
- What metrics to report?
- How to interpret external validation?



Guideline > Ann Intern Med. 2015 Jan 6;162(1):55-63. doi: 10.7326/M14-0697.

Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement

Gary S Collins, Johannes B Reitsma, Douglas G Altman, Karel G M Moons

Improving visual communication of discriminative accuracy for predictive models: the probability threshold plot

Stephen S Johnston, Stephen Fortin, Iftekhar Kalsekar, Jenna Reps, Paul Coplan

JAMIA Open, Volume 4, Issue 1, January 2021, ooab017,
https://doi.org/10.1093/jamiaopen/ooab017

Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation

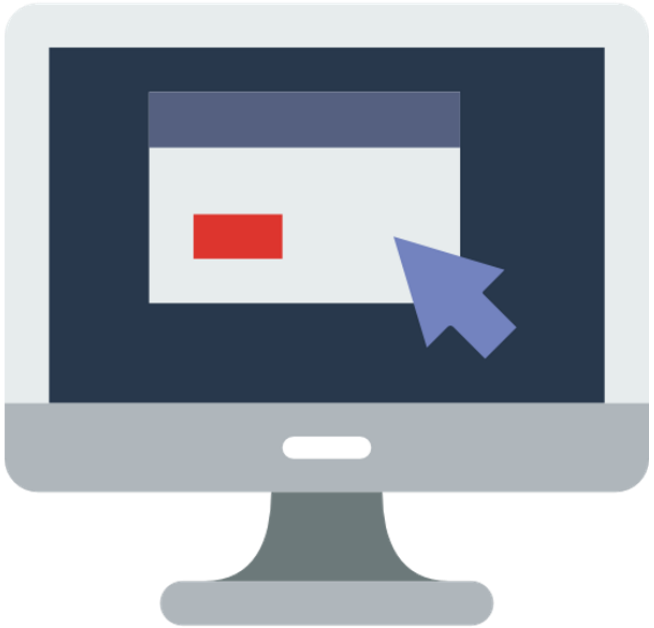
Jenna M. Reps, Ross D. Williams, Seng Chan You, Thomas Falconer, Evan Minty, Alison Callahan, Patrick B. Ryan, Rae Woong Park, Hong-Seok Lim & Peter Rijnbeek

BMC Medical Research Methodology **20**, Article number: 102 (2020) | Cite this article

The process of using the model

Model Deployment

- How to implement the model?
- Are there constraints?



Model Deployment

A framework for making predictive models useful in practice

Kenneth Jung, Sehj Kashyap, Anand Avati, Stephanie Harman, Heather Shaw, Ron Li, Margaret Smith, Kenny Shum, Jacob Javitz, Yohan Vetteth ... Show more

Journal of the American Medical Informatics Association, Volume 28, Issue 6, June 2021, Pages 1149–1158, https://doi.org/10.1093/jamia/ocaa318

Published: 22 December 2020 **Article history ▾**

Take Home

- Machine learning could be applied to develop useful prediction models– including for **causal inference**
- But we need to follow best practices
- More work is required to get empirical evidence that supports best practices for using **machine learning** to develop prediction models using observational data

Thanks for listening

- Email: jreps@its.jnj.com

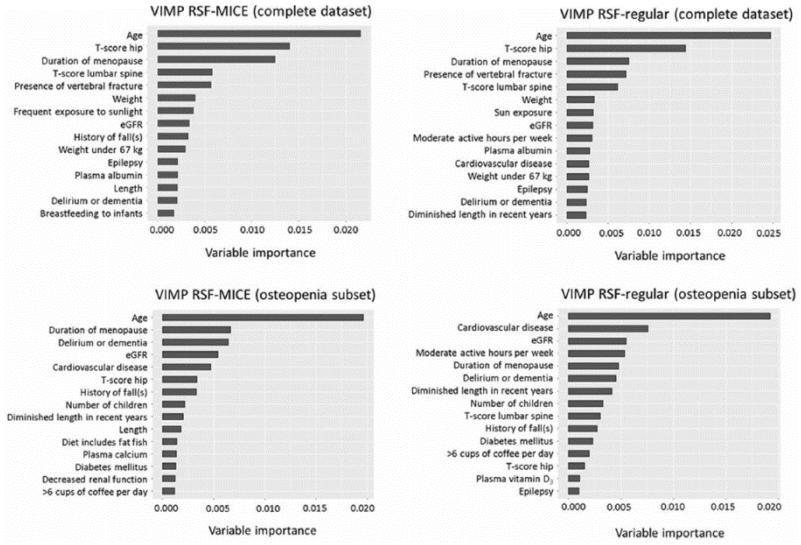
Karin Groothuis-Oudshoorn, PhD
University of Twente
Enschede, Netherlands

Virtual

ISPOR Europe 2021 Spotlight 1

In "Deep" Thought: Advancing Machine Learning Methods in HEOR

Deep though: advancing machine learning methods in HEOR



Discussant:

Karin Groothuis-Oudshoorn

UNIVERSITY
OF TWENTE.

ML for causal inference

- Prediction is easier than attribution or estimation (Efron, 2020)
  - Pure prediction algorithms operate non parametrically, do not worry on estimation efficiency
  - Weak learners are not available for attribution
- Biostatistical models can help with all 3 tasks
  - Explainability is not an issue for these models
- But even prediction turns out difficult in a lot of cases

Efron, B. Prediction, Estimation, and Attribution, JASA, 2020:115(530):636-655

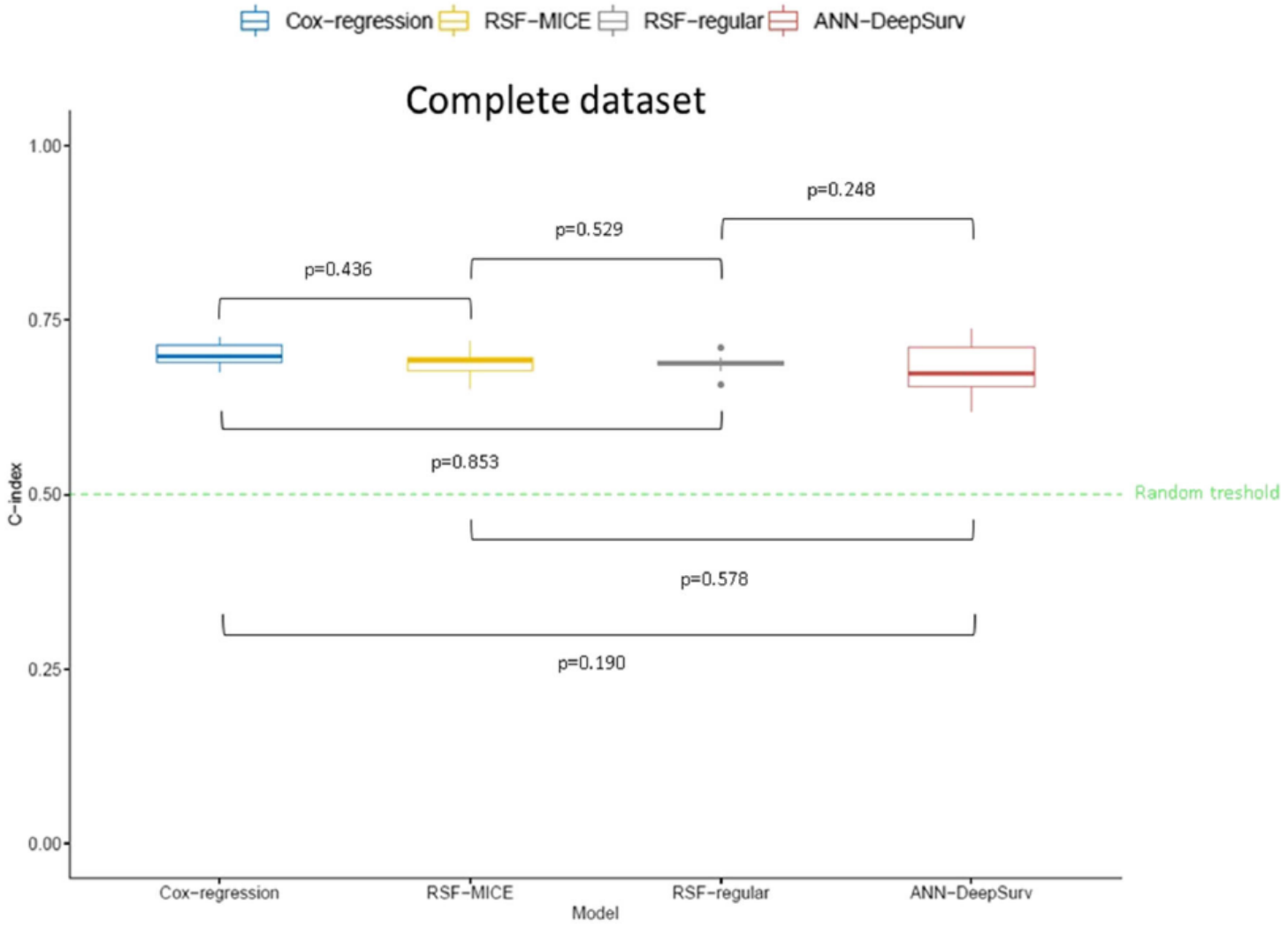
Challenges use of prediction models in clinical practice

- Most studies on ML based prediction models show poor methodological quality & high risk of bias (Navarro, 2021)
- When offers ML really an improvement?
  - Performance
  - Clinical utility

Navarro, C et all. Risk of bias in studies on prediction models developed using supervised machine learning techniques: a systematic review, BMJ, 2021; 375

Predicting a subsequent Major Osteoporotic Fracture

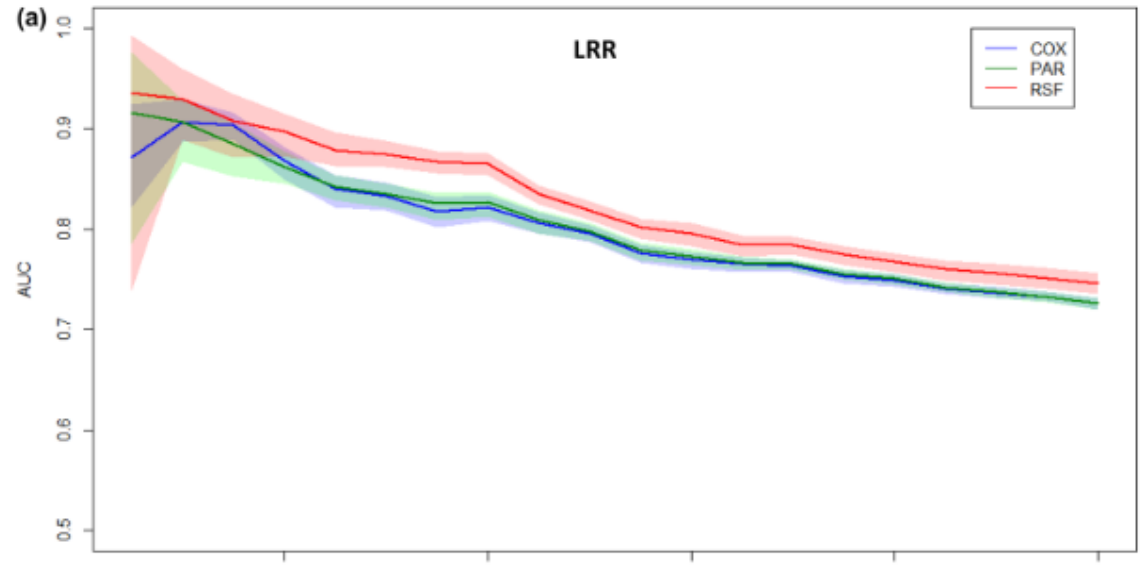
- Data from EHR
- Time till occurrence of a MOF was extracted using diagnosis treatment combinations (DBCs)
- Starting date: visit to the Fracture and Osteoporosis outpatient clinic
- 7578 patients, 805 with a subsequent MOF
- Comparison of different methods:
  - Cox regression model
  - Random Survival Forest
  - Artificial Neural Network - DeepSurv model
- Validation with 10-fold crossvalidation

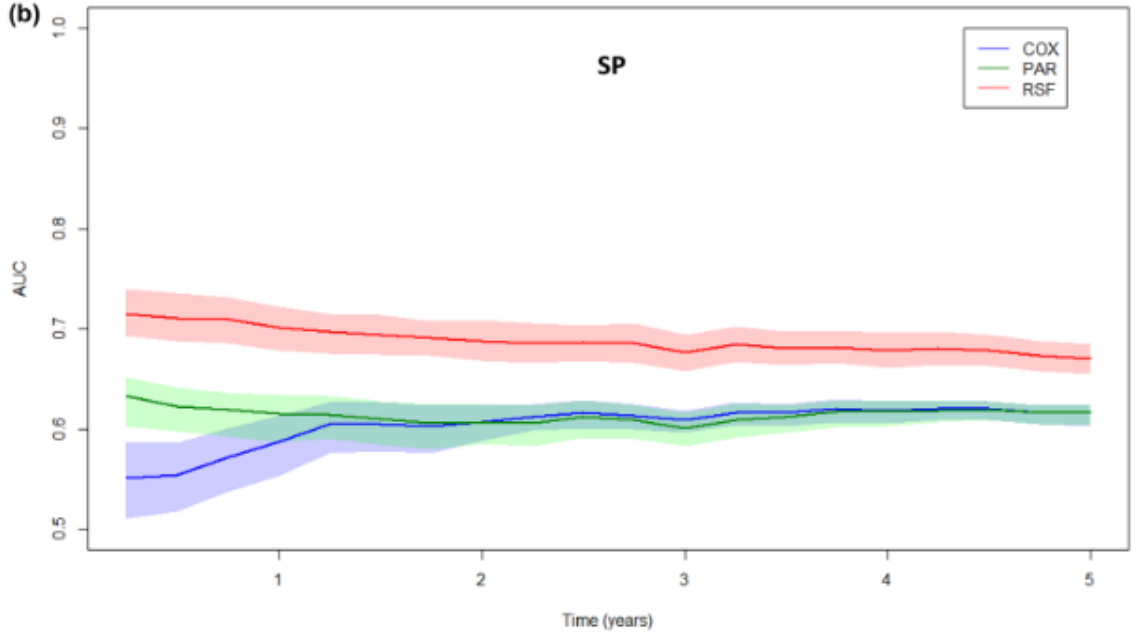


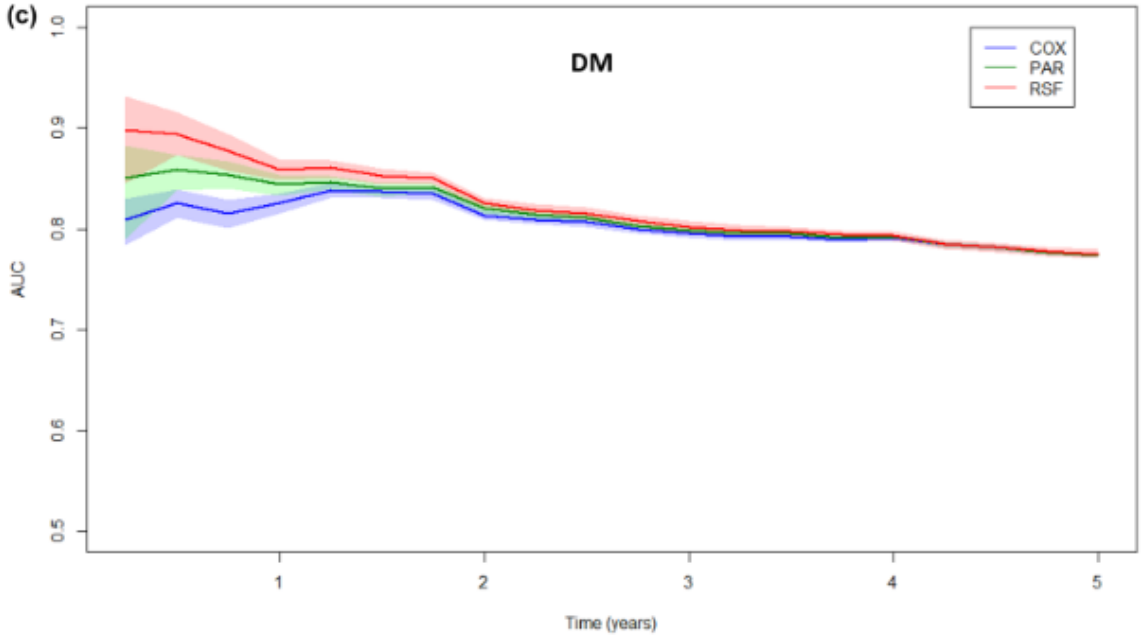
Vries BCS, Hegeman J, Nijmeijer W, Geerdink J, Seifert C, Groothuis-Oudshoorn C. Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis, Osteoporosis International, 2020

INFLUENCE 2.0 for risk prediction of breast cancer patients

- Data from the Dutch Cancer Registry
- Time dependent individual risk prediction for LRR, SP and DM (up to 5 years)
- 13494 patients included
  - 2.8% developed LRR
  - 3.0% developed SP
  - 6.3% developed DM
- Comparison of different methods:
- Cox regression model
- Parametric spline approach
- Random Survival forest
- 11 covariates
- Validation with bootstrap validation (Harrell's method)







Völkel V, Hueting T, Draeger T, van Maaren M, Munck L, Strobbe L, Sonke G, Schmidt M, Hezewijk M, Siesling S, Groothuis-Oudshoorn C. Improved risk estimation of locoregional recurrence, secondary contralateral tumors and distant metastases in early breast cancer: the INFLUENCE 2.0 model, Breast Cancer Research and Treatment (2021) 189:817-826

Calibration

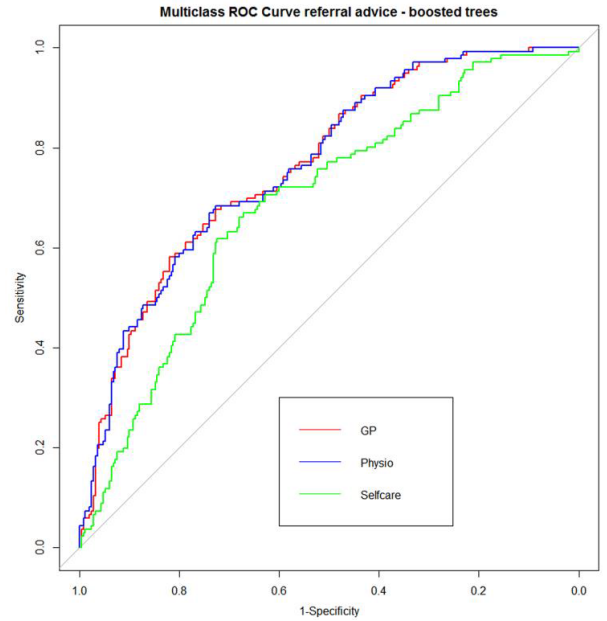
Table 2 Calibration results

Outcome	Time (years)	COX			PAR			RSF		
		ICI	E50	E90	ICI	E50	E90	ICI	E50	E90
Locoregional	1	0.0005	0.0003	0.0007	0.0019	0.0014	0.0016	**0.0009**	**0.0006**	**0.0015**
	2	0.0011	0.0008	0.0016	0.0030	0.0021	0.0025	**0.0023**	**0.0018**	**0.0033**
	3	0.0017	0.0013	0.0027	0.0028	0.0022	0.0032	**0.0044**	**0.0023**	**0.0039**
	4	0.0023	0.0018	0.0037	0.0023	0.0019	0.0036	**0.0064**	**0.0035**	**0.0055**
	5	0.0027	0.0022	0.0044	0.0028	0.0021	0.0042	**0.0094**	**0.0057**	**0.0096**
Second primary	1	0.0010	0.0009	0.0017	0.0030	0.0024	0.0060	**0.0010**	**0.0008**	**0.0017**
	2	0.0013	0.0011	0.0023	0.0020	0.0016	0.0036	**0.0018**	**0.0014**	**0.0032**
	3	0.0016	0.0015	0.003	0.0022	0.0021	0.0033	**0.0037**	**0.0028**	**0.0063**
	4	0.0021	0.0018	0.0038	0.0021	0.0019	0.0036	**0.0049**	**0.004**	**0.0088**
	5	0.0025	0.0022	0.0044	0.0026	0.0022	0.0047	**0.0073**	**0.0059**	**0.0135**
Distant metastasis	1	**0.0007**	**0.0004**	**0.0012**	0.0030	0.0024	0.0034	0.0014	0.0008	0.0017
	2	**0.0017**	**0.0009**	**0.0024**	0.0055	0.0035	0.0062	0.0060	0.0036	0.0058
	3	**0.0026**	**0.0015**	**0.0038**	0.0054	0.0035	0.0075	0.0119	0.0074	0.0123
	4	**0.0032**	**0.0020**	**0.0045**	0.0043	0.0028	0.0075	0.0166	0.0106	0.0201
	5	**0.0037**	**0.0024**	**0.0054**	0.0033	0.0024	0.0054	0.0216	0.0143	0.0299

The displayed results represent the optimism-corrected values. The ICI is the integrated calibration index, E50 is the median absolute difference between observed and expected, and E90 is the 90th percentile of the absolute difference. Bold results display the results for the models that were selected as the best-performing model. For LRR, and SP, the decision was primarily based on the discrimination

Issues in performance of prediction models

- When the data availability is the same, more ‘advanced’ ML models will not help?
- How to disentangle outcome and treatment decision
- The number of cases is a hampering factor
- Calibration is also important
- ROC curve with alternative validation methods (bootstrap / CV)?
- Multiclass prediction is much harder compared to binary classification, but might be more relevant



ELSEVIER

International Journal of Medical Informatics
Volume 110, February 2018, Pages 31-41

Evaluation of three machine learning models for self-referral decision support on low back pain in primary care

Wendy Oude Nijeweme-d'Hollosy [a], Lex van Velsen [a, b], Mannes Poel [c], Catharina G.M. Groothuis-Oudshoorn [d], Remko Soer [e, f], Hermie Hermens [a, b]

Show more

What's next?

- Clinical vignette studies to capture decision process clinicians
- Preferences on what data are patients willing to share
  - Include ethical values into development ML models
- Transferability of prediction models
  - Standard of care is confounded with the data
  - www.evidencio.nl: database of >2000 medical algorithms
- Integration and use of different type of data
  - Images, videos
  - Longitudinal data with different time scales
  - Heterogeneity in features
- New certification regulations (EU MDR)

CE FDA

ISPOR Europe 2021 Spotlight 1

In "Deep" Thought: Advancing Machine Learning Methods in HEOR

William Crown, PhD
Brandeis University
Waltham, MA, USA

Vivek Rudrapatna, MD, PhD
University of California
School of Medicine
San Francisco, CA, USA

Jenna Reps, PhD
Janssen Pharmaceuticals, Inc.
Raritan, NJ, USA

Karin Groothuis-Oudshoorn, PhD
University of Twente
Enschede, Netherlands