Welcome to the Sentinel Innovation and Methods Seminar Series

The webinar will begin momentarily

Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Sentinel

Deep Learning on Electronic Health Records for Research in Pharmacoepidemiology: Examples From the Field of Oncology

Janick Weberpals, RPh, PhD

Sentinel Innovation Center Webinar
September 7, 2022

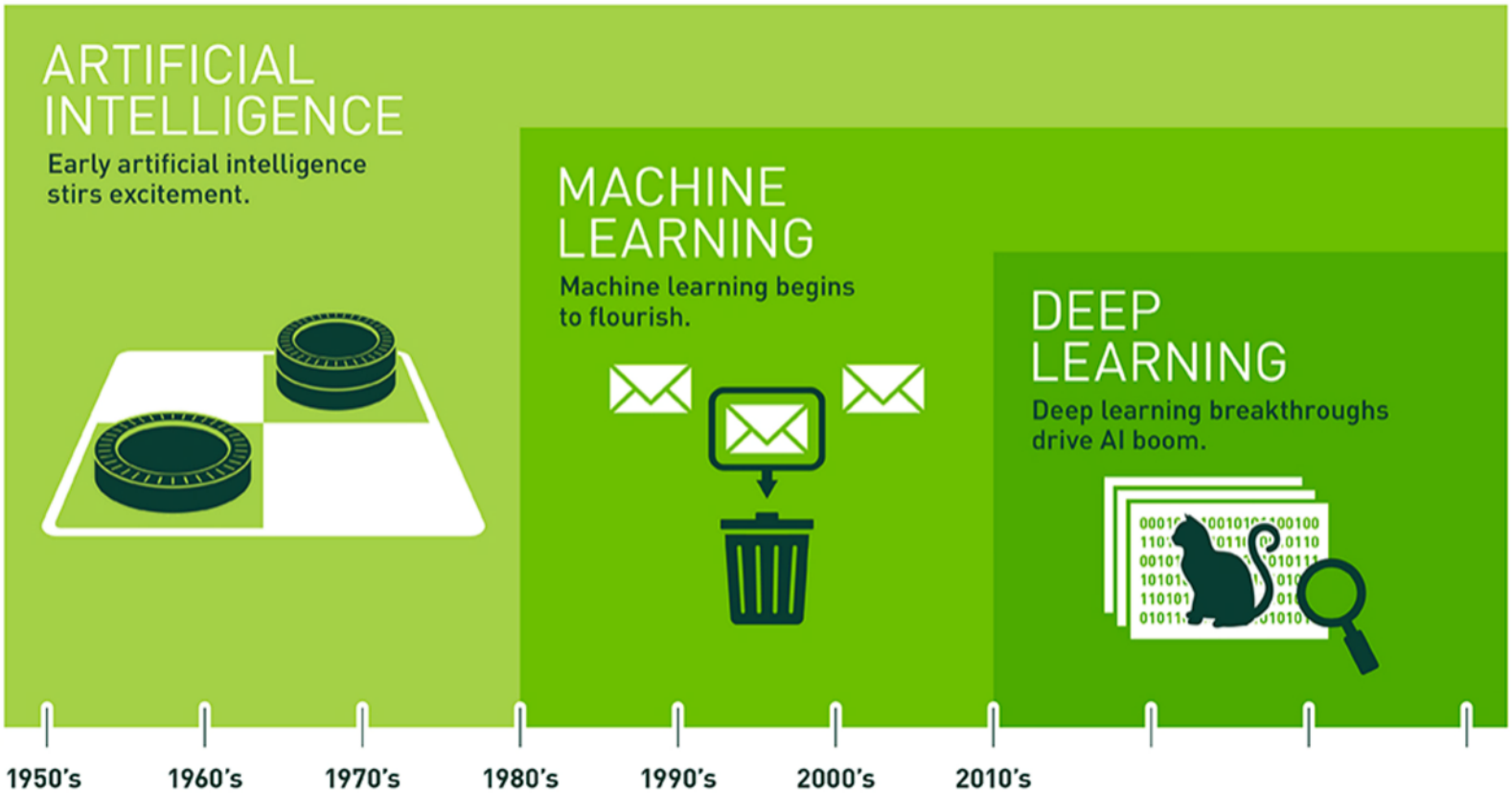
jweberpals@bwh.harvard.edu

janickweberpals.github.io

Division of Pharmacoepidemiology and Pharmacoeconomics
Department of Medicine, Brigham and Women's Hospital and Harvard Medical School

Disclosures

- The presenter, Janick Weberpals, is a former employee of Hoffmann-La Roche Ltd. and held shares in Hoffmann-La Roche Ltd.
- The research in this presentation was conducted during the presenters postdoctoral fellowship program funded by Hoffmann-La Roche Ltd.
- All studies covered in this presentation are published and are publicly accessible



Computational approaches that aim to mimic (human) intelligence

Make predictions from data by "learning by examples"

A type of machine learning using highly flexible, complex algorithms

Adapted from: https://towardsdatascience.com/deep-learning-weekly-piece-the-differences-between-ai-ml-and-dl-b6a203b70698

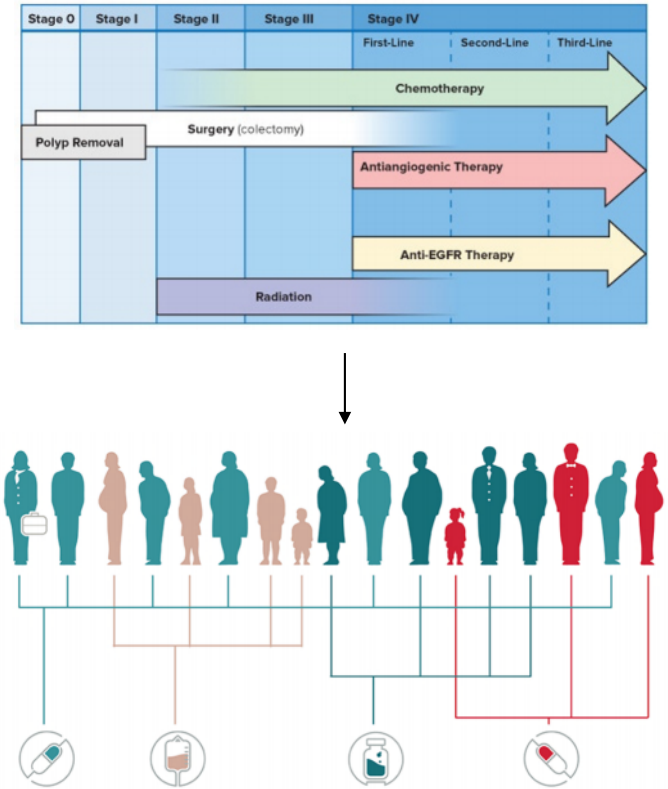
Can deep learning significantly enhance our ability to make **causal predictions** in comparative effectiveness and safety research?

Prognostic/predictive scores in oncology

- Clinical decision making
  - Treatment decision making
  - NCCN guidelines partly rely on risk models
  - Trial eligibility criteria (e.g. expected survival)
- Clinical drug development and basic research
- Methodologically interesting (disease risk scores)
- Patient need for information about the future

➢ Historically TNM staging used to be most important information

➢ Changes in era of precision oncology
  - o Biomarker
  - o Digital pathology
  - o Tumor-agnostic approvals (e.g. MSI-high tumors)
  - o Multimodal prognostic/predictive scores



Targeted therapy for subgroups of patients selected via the right diagnostic tools or biomarkers.

Motivation

Contemporary prognostic models

Royal Marsden Hospital Hospital Score (RMHS), International Prognostic Index (IPI), Glasgow prognostic score (GPS)

+

Increasing access to more data

Horizontal information (more variables)
Vertical information (larger sample size)

+

Increasing computational resources

Advanced analytical methods such as deep learning

Increased prognostic performance?

FDA provided updated guidance on patient enrichment strategies in investigational studies[1] aiming to

(i) decrease interpatient variability
(ii) identify high-risk patients to enable prognostic enrichment strategies
(iii) to identify more responsive patients for predictive enrichment

[1] **Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products (March 2019).** Available at: http://www.fda.gov/regulatory-information/search-fda-guidance-documents/enrichment-strategies-clinical-trials-support-approval-human-drugs-and-biological-products

Real wOrld PROgnostic score (ROPRO)

ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY driving innovation in oncology

ORIGINAL ARTICLE

An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort

T. Becker[1], J. Weberpals[1], A. M. Jegg[2], W. V. So[3], A. Fischer[1], M. Weisser[2], F. Schmich[1], D. Rüttinger[2†] & A. Bauer-Mehren[1*†]

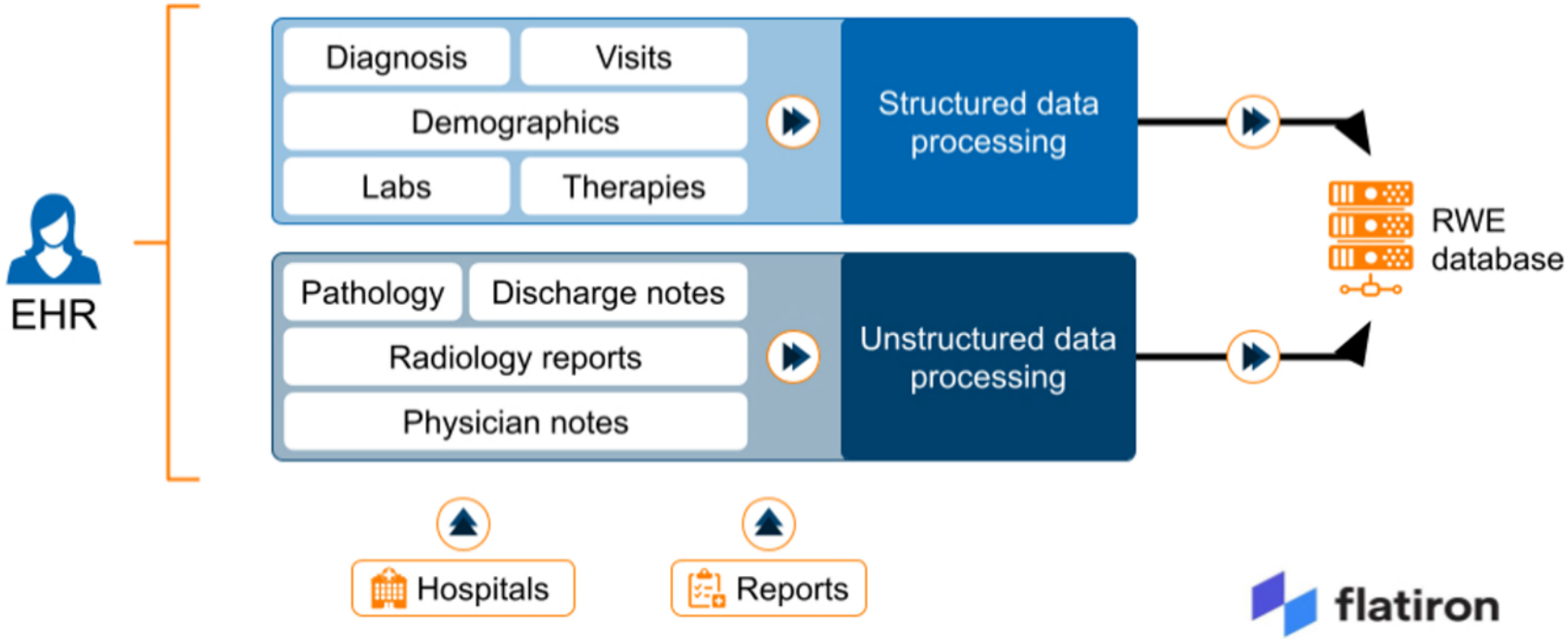
[1]Data Science, Pharma Research and Development, Roche Innovation Center Munich, Munich; [2]Early Clinical Development Oncology, Pharma Research and Development, Roche Innovation Center Munich, Munich, Germany; [3]Data Science, Pharma Research and Development, Roche Innovation Center New York, New York, USA

CHECK FOR UPDATES

Available online 31 July 2020

Becker T, Weberpals J, Jegg AM, So WV, Fischer A, Weisser M, Schmich F, Rüttinger D, Bauer-Mehren A. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. Ann Oncol. 2020 Nov;31(11):1561-1568.

Methods - Database and covariate ascertainment



EHR: electronic health records; RWE: real-world evidence.

Study design (adapted from Schneeweiss S. et al. Ann Intern Med. 2019 Mar 19;170(6):398-406)

I/E criteria {T0}:
- Diagnosed with one of the 17 available cancer types in Flatiron Health database
- Any 1L therapy for respective cancer
- High missingness in records (>75%)

I/E criteria {-30 days;0}:
- Labs
- Biomarker
- Vital status
- ...

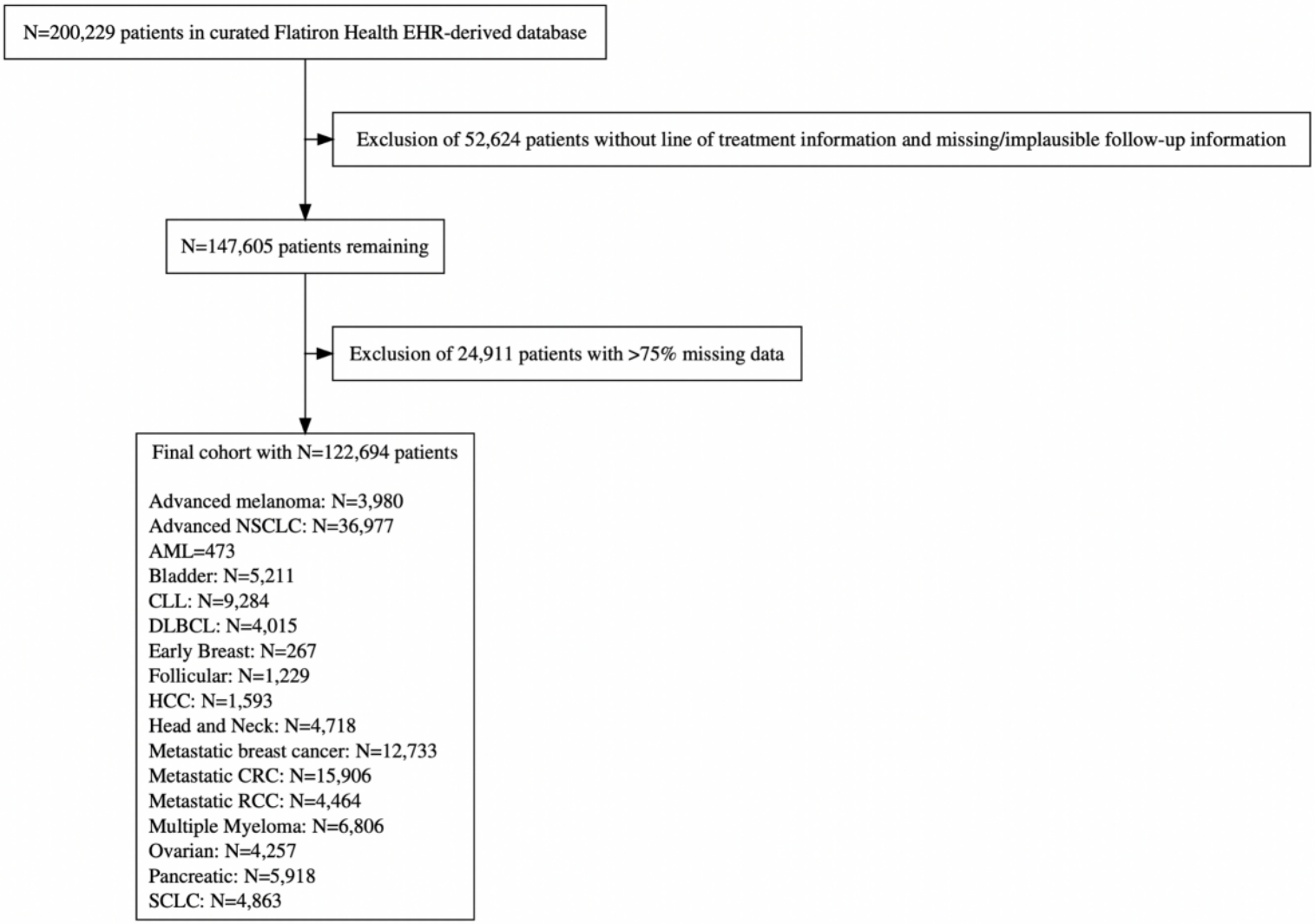
I/E criteria {T0}:
- Age
- Gender
- ...

Follow-up time was computed as time from index date to the date of death due to any reason or to patient's last documented clinic contact or database lock (31 December 2019 | database version Feb 2020), whichever occurred first

Follow-up

T0 / index date
(initiation of 1L therapy for respective cancer type)

Becker T, Weberpals J, Jegg AM, So WV, Fischer A, Weisser M, Schmich F, Rüttinger D, Bauer-Mehren A. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. Ann Oncol. 2020 Nov;31(11):1561-1568.


N=200,229 patients in curated Flatiron Health EHR-derived database
Exclusion of 52,624 patients without line of treatment information and missing/implausible follow-up information
N=147,605 patients remaining
Exclusion of 24,911 patients with >75% missing data
Final cohort with N=122,694 patients
Advanced melanoma: N=3,980
Advanced NSCLC: N=36,977
AML=473
Bladder: N=5,211
CLL: N=9,284
DLBCL: N=4,015
Early Breast: N=267
Follicular: N=1,229
HCC: N=1,593
Head and Neck: N=4,718
Metastatic breast cancer: N=12,733
Metastatic CRC: N=15,906
Metastatic RCC: N=4,464
Multiple Myeloma: N=6,806
Ovarian: N=4,257
Pancreatic: N=5,918
SCLC: N=4,863

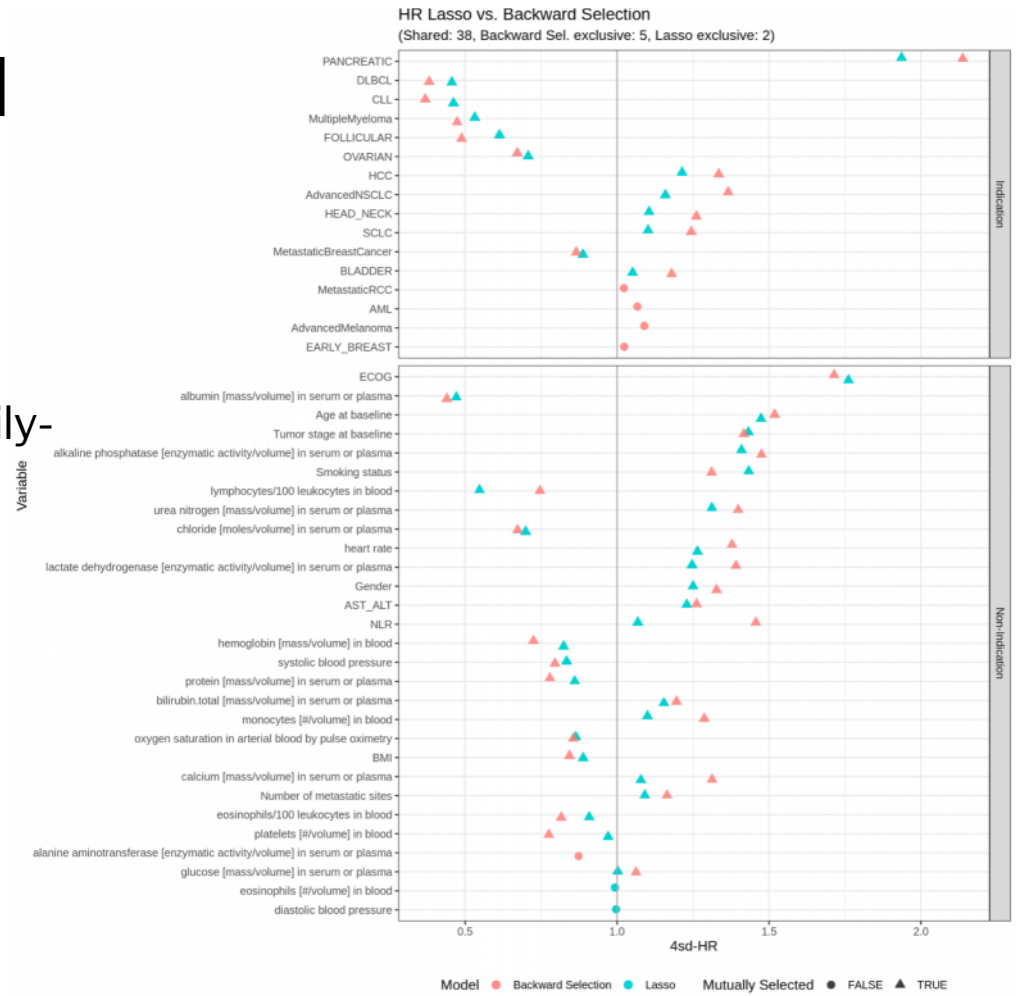
Covariate selection and modelling

- Model selection on 46 variables
- Main model: Backward selection with family-wise error rate
- COX-LASSO plus 10-fold cross-validation

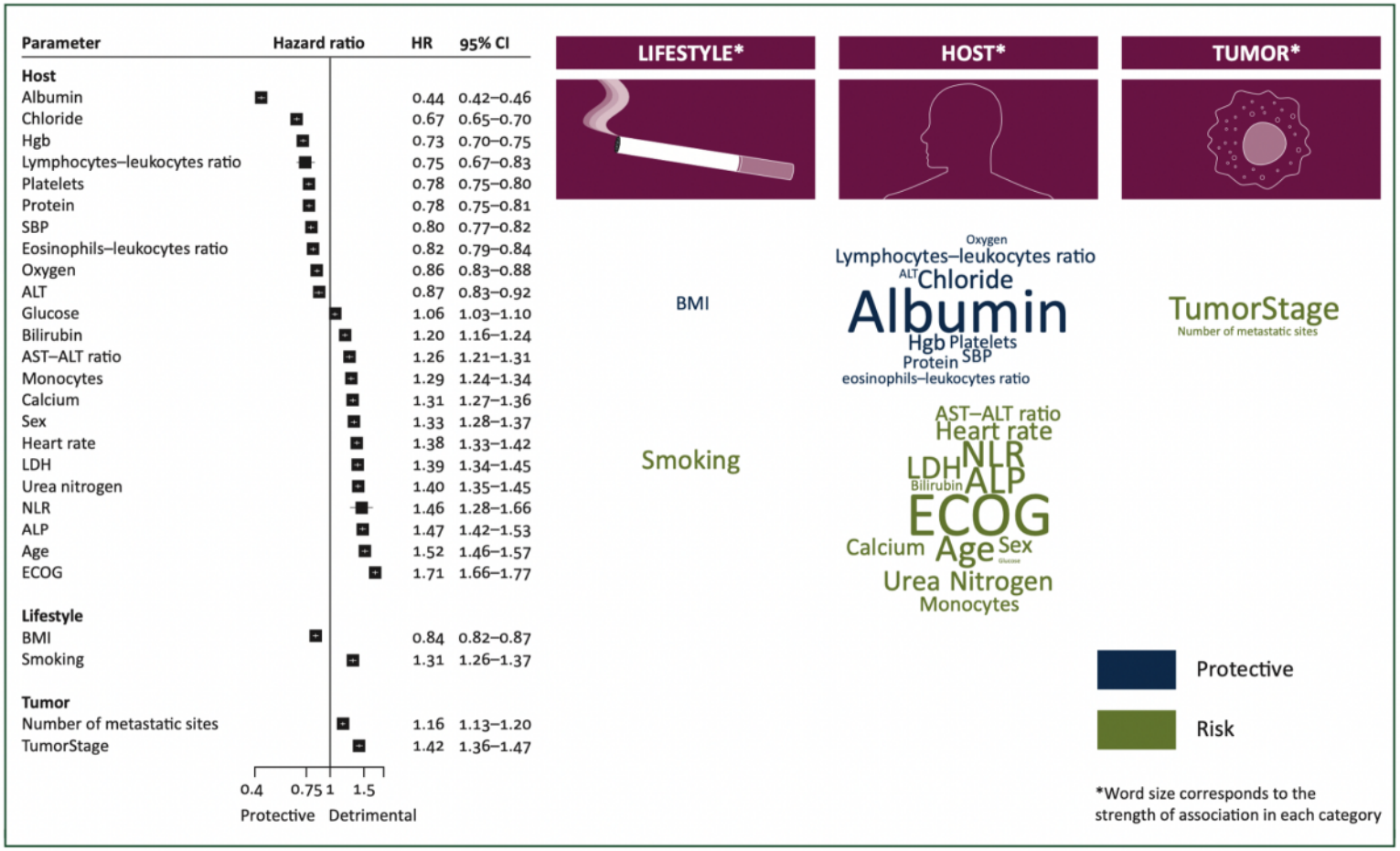
ROPRO was specified as:

$$\sum_i \ln(HR_{xi})(m_{ij} - m_i)$$

HR_{xi} : estimated HR for variable i
m_i : variable mean in Flatiron Health dataset
m_{ij} : variable value of patient j for variable i



Results



Model selection resulted in highly prognostic variables:

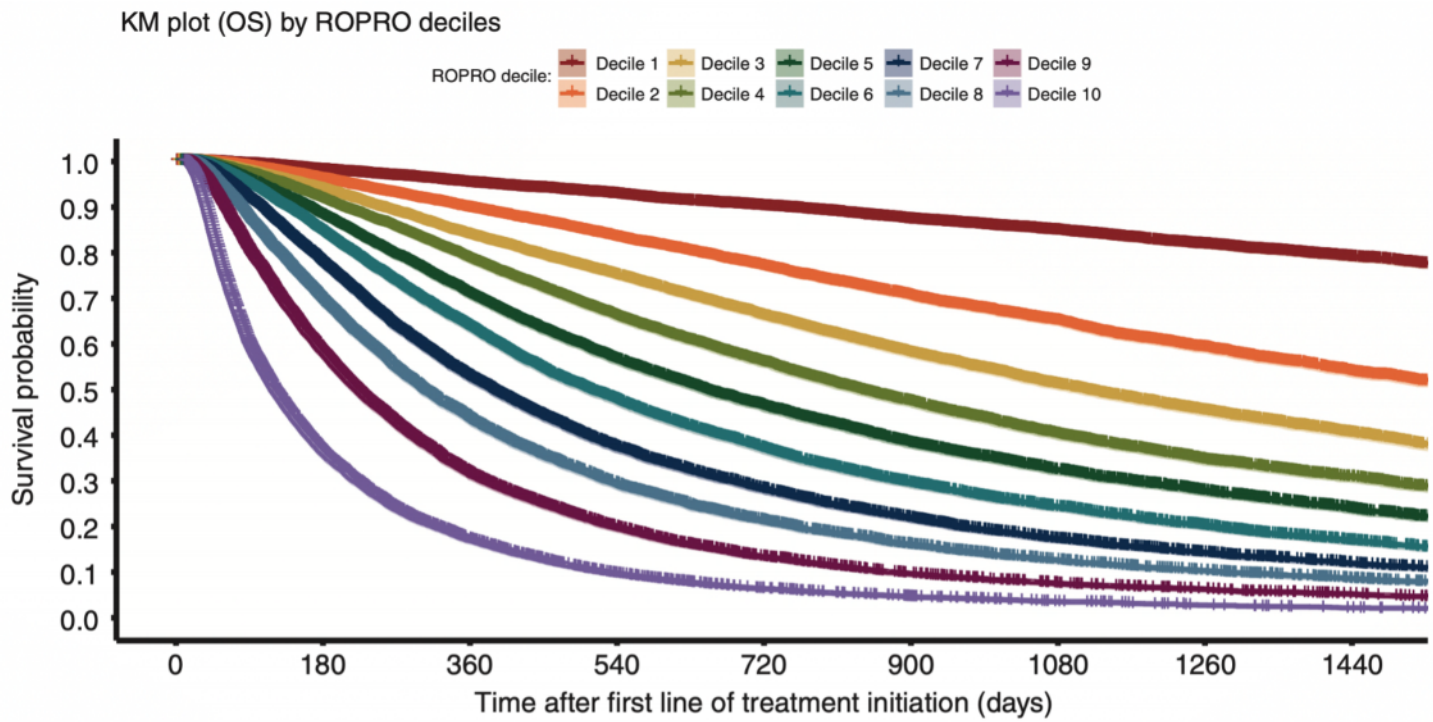
- **27 covariates in backward selection**
- 28 covariates in LASSO model
- 26 covariates coincided (Pearson correlation of r^2 = 0.993)

Resulting patient individual **ROPRO score** is based on a weighted sum of the patients' differences from the respective reference means of each variable (according to formula shown earlier)

↑ROPRO score = ↑$Hazard_{OS}$

Becker T, Weberpals J, Jegg AM, So WV, Fischer A, Weisser M, Schmich F, Rüttinger D, Bauer-Mehren A. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. Ann Oncol. 2020 Nov;31(11):1561-

Results



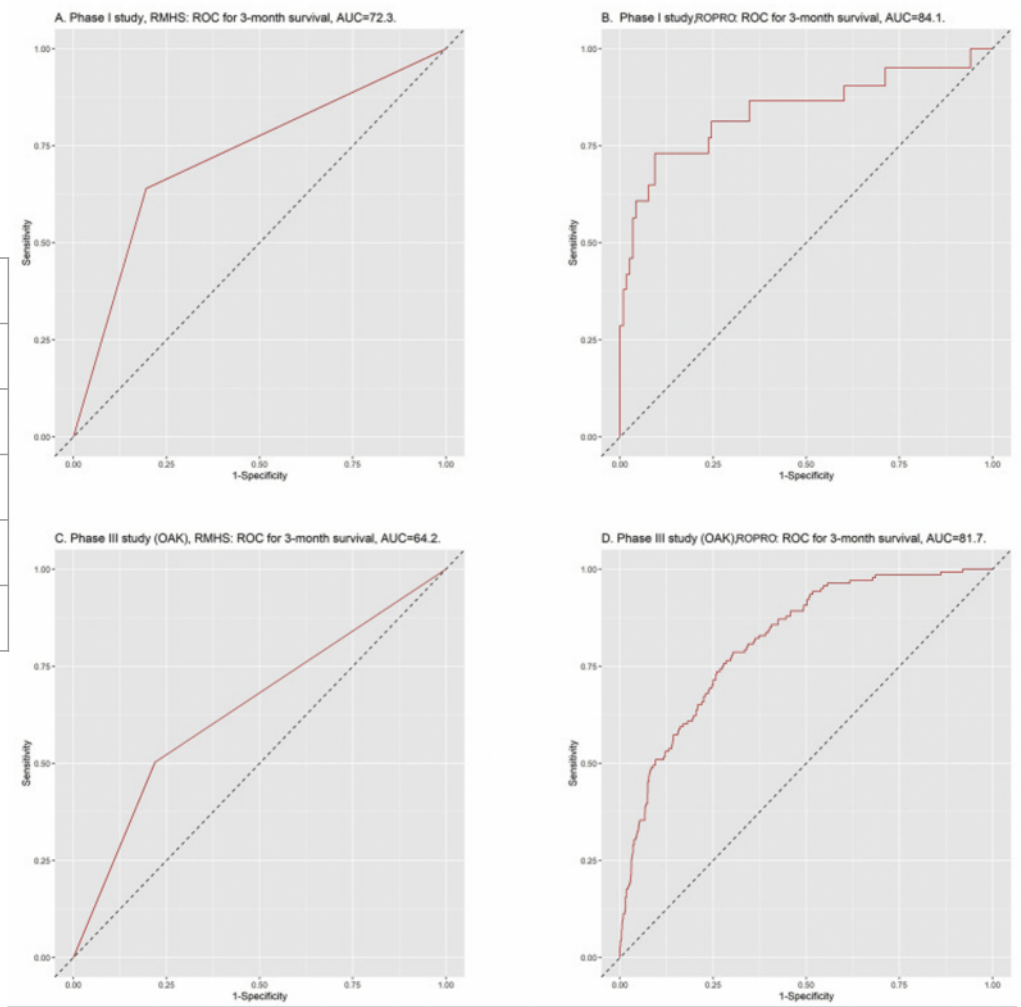
Number at risk

ROPRO decile:	0	180	360	540	720	900	1080	1260	1440
Decile 1	12 270	10 885	9629	8502	7490	6467	5543	4721	3964
Decile 2	12 269	10 600	8984	7481	6204	5027	4084	3212	2472
Decile 3	12 269	10 379	8321	6648	5227	4027	3089	2382	1775
Decile 4	12 270	10 137	7810	5831	4363	3215	2376	1731	1299
Decile 5	12 269	9752	7029	4985	3603	2599	1910	1376	980
Decile 6	12 269	9309	6313	4177	2882	1996	1434	1015	691
Decile 7	12 270	8512	5151	3247	2158	1457	1006	688	478
Decile 8	12 269	7495	4131	2451	1535	1025	686	491	322
Decile 9	12 269	6169	2947	1613	946	602	392	278	185
Decile 10	12 270	3751	1556	767	426	252	157	101	65

Time after first line of treatment initiation (days)

Results

ROPRO versus RMHS in development dataset (Flatiron Health)		
Metrics	**ROPRO (pan-tumor)**	**RMHS**
Generalized R^2	0.319	0.033
C-index (sd)	0.747 (0.0012)	0.541 (0.0005)
AUC $_{\text{3-month survival}}$	0.822	0.579
AUC $_{\text{1-year survival}}$	0.804	0.549



Becker T, Weberpals J, Jegg AM, So WV, Fischer A, Weisser M, Schmich F, Rüttinger D, Bauer-Mehren A. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. Ann Oncol. 2020 Nov;31(11):1561-

Can we boost prognostic performance using more complex ML/DL models?

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 16 April 2021
doi: 10.3389/frai.2021.625573

Artificial Intelligence for Prognostic Scores in Oncology: a Benchmarking Study

Hugo Loureiro[1,2,3], Tim Becker[1], Anna Bauer-Mehren[1], Narges Ahmidi[2†] and Janick Weberpals[1†]*

[1]Data Science, Pharmaceutical Research and Early Development Informatics (pREDi), Roche Innovation Center Munich (RICM), Penzberg, Germany, [2]Institute of Computational Biology, Helmholtz Zentrum Munich, Munich, Germany, [3]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany

Loureiro H, Becker T, Bauer-Mehren A, Ahmidi N, Weberpals J. Artificial Intelligence for Prognostic Scores in Oncology: a Benchmarking Study. Front Artif Intell. 2021 Apr 16;4:625573.

Study setup

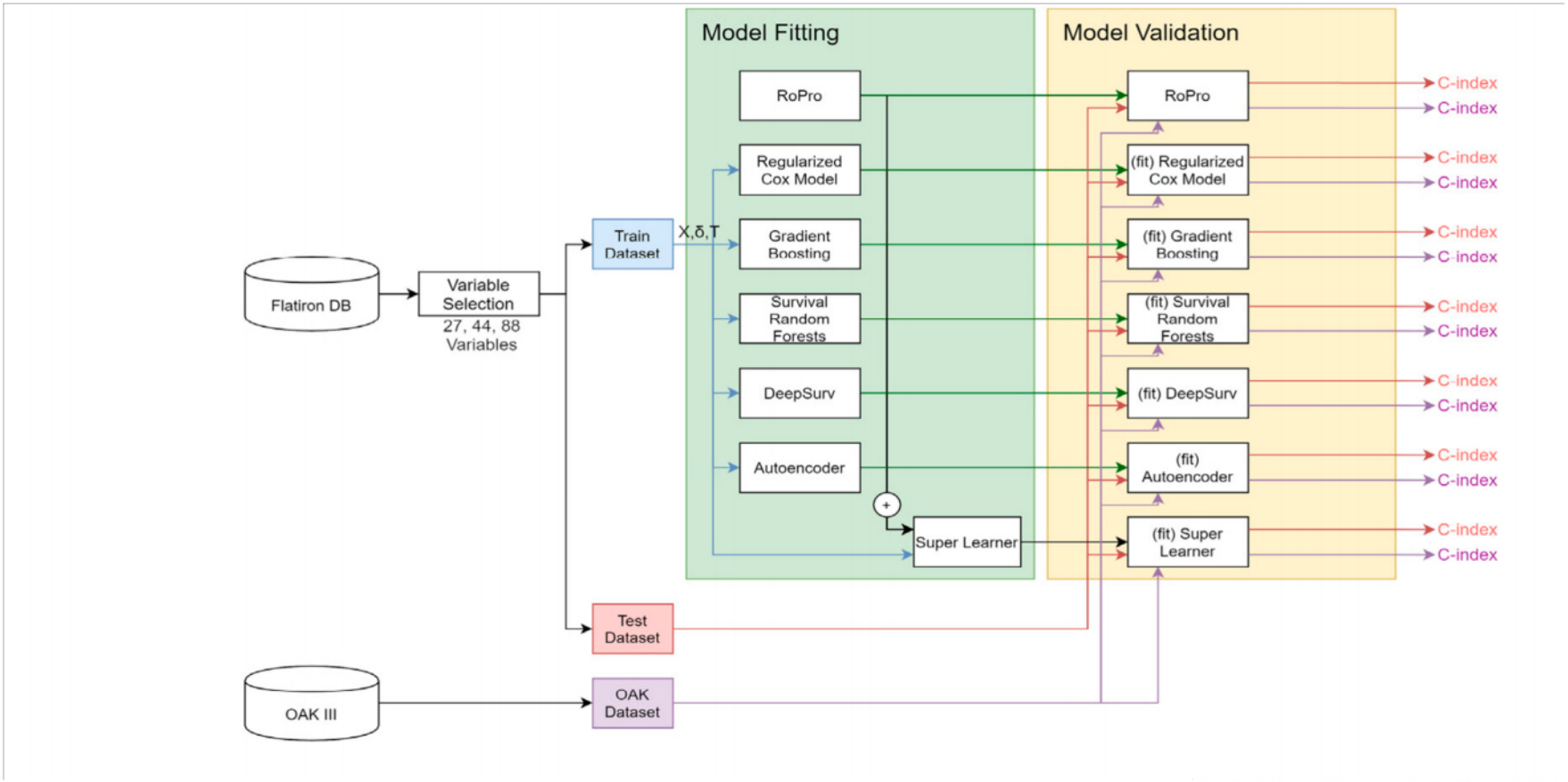
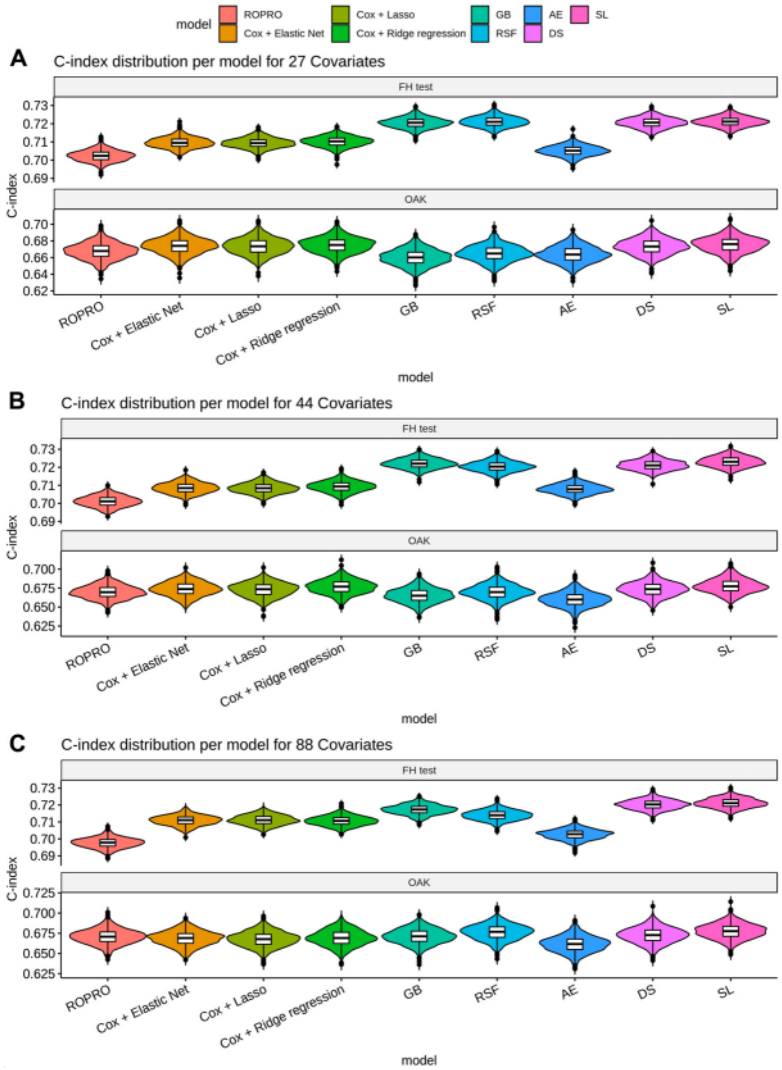


FIGURE 1 | Diagram of the analysis.

Loureiro H, Becker T, Bauer-Mehren A, Ahmidi N, Weberpals J. Artificial Intelligence for Prognostic Scores in Oncology: a Benchmarking Study. Front Artif Intell. 2021 Apr 16;4:625573.

Frontiers in artificial intelligence 4 (2021): 9.



Results

- Similar patterns across all covariate sets
- In 44 covariate **FH test** set:
  - ROPRO benchmark C-index 0.701 [0.696, 0.706]
  - Model performances meaningfully improved using SL (C-index 0.723 [0.718, 0.728])
- In 44 covariate **OAK validation** set:
  - Model that yielded the highest C-index was SL 0.677 [0.662, 0.695] vs ROPRO 0.670 [0.657, 0.685]
  - Meaningful improvement of more complex model in FH test set disappeared

Interpretation

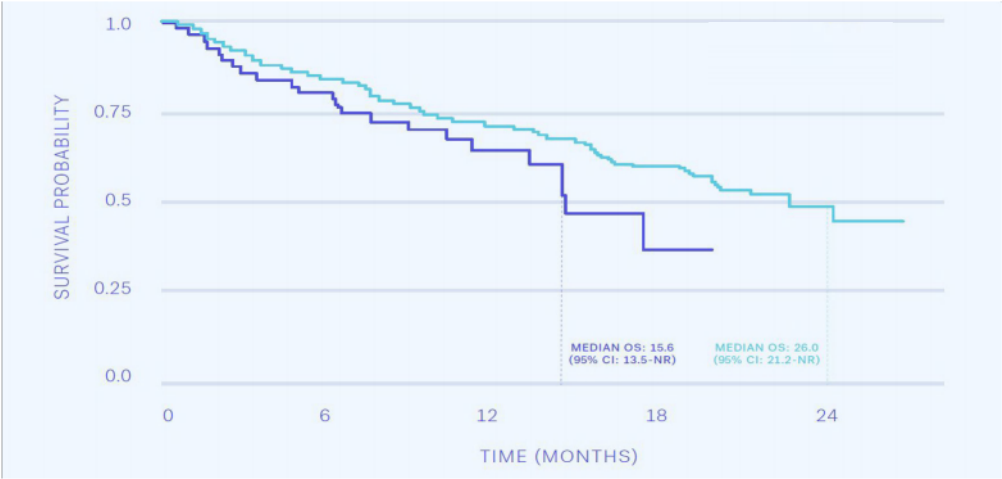
- Conclusion: using complex machine learning models did not meaningfully increase the performance of prognostic scores in oncology
- Similar observations also made in other domains (e.g. HF, *Desai RJ et al., JAMA Netw Open 2020*)
- Covariates used for prediction rather limited

Loureiro H, Becker T, Bauer-Mehren A, Ahmidi N, Weberpals J. Artificial Intelligence for Prognostic Scores in Oncology: a Benchmarking Study. Front Artif Intell. 2021 Apr 16;4:625573.

Can deep learning significantly enhance our ability to make **causal inference** in comparative effectiveness and safety research?

Comparative effectiveness studies

Drug A



Drug B

- There might be systematic differences in baseline characteristics between patients who received Drug A vs. B
- Use of **propensity scores**
  - Conditional probability that an individual receives a certain treatment based on baseline characteristics

$$Pr(Z_i = 1|X_i)$$

  - <u>Theory</u>: by conditioning (matching, weighting, …) the two cohorts on the propensity score the only difference is treatment

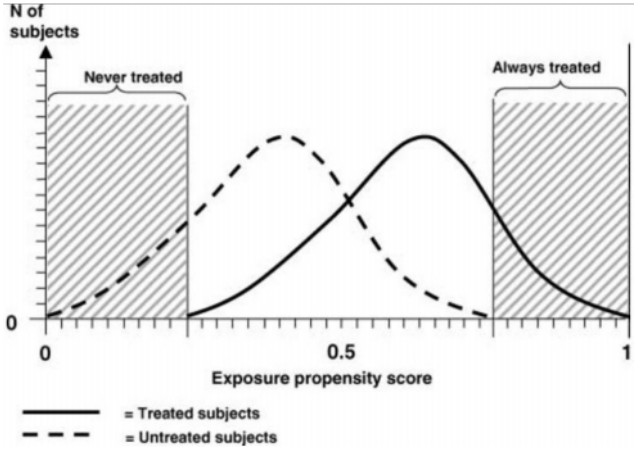
Image: Davies J. et al., J Comp Eff Res (2018)

Propensity scores - assumptions

The validity of results derived through propensity score analysis comes with assumptions

1. **No unmeasured confounding**
   - Often difficult to assess & test
   - Potential solutions: high-dimensional propensity scores, IV analysis, active comparator designs, …

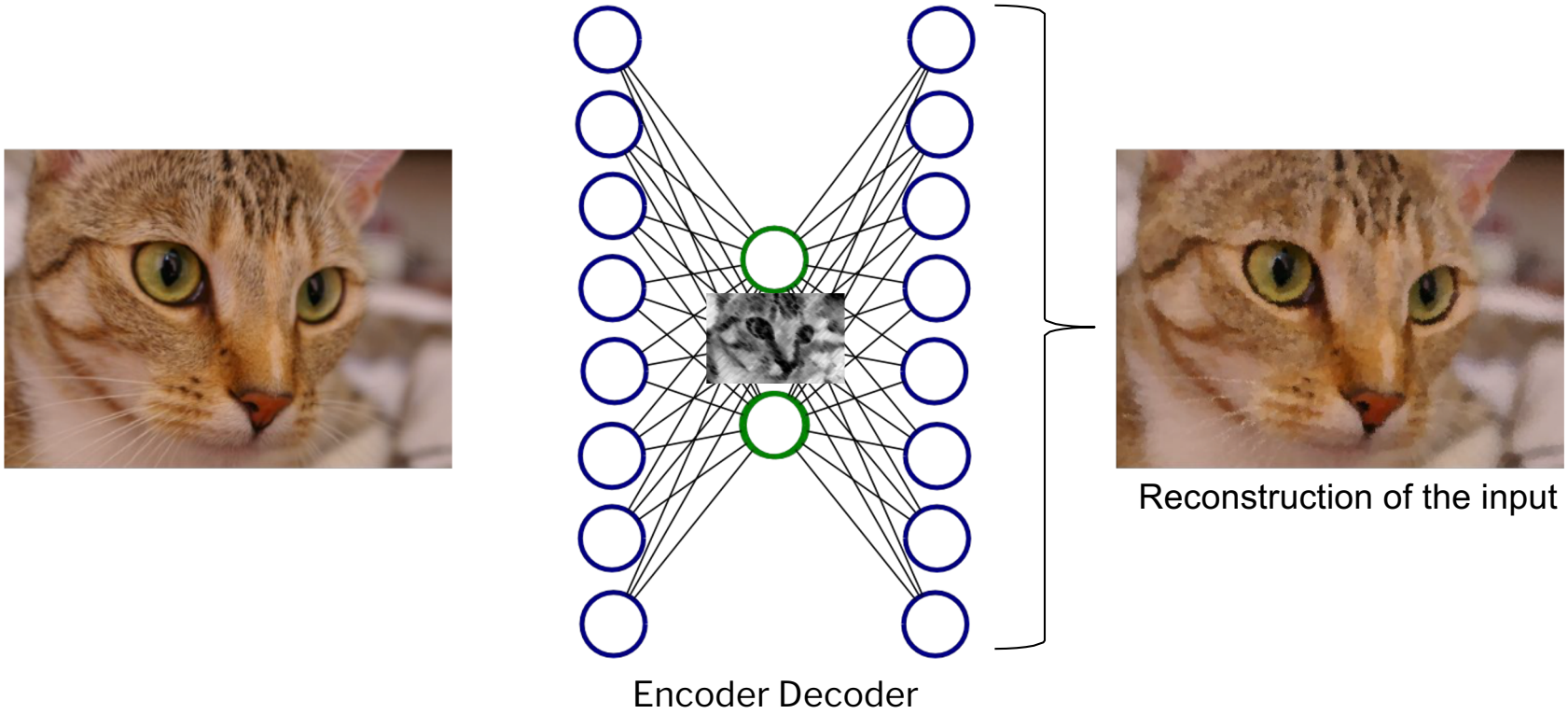
2. **Propensity score model has to be correctly specified**
   - Variable selection
     - Logistic regression fitted using a-priori investigator defined variables (Literature, expert knowledge, …)
     - Predictors of treatment <u>and</u> outcome
     - Predictors of outcome
   - Non-linearities & non-additivities often not considered



Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-1156. doi:10.1093/aje/kwj149
RJ Glynn, S Schneeweiss, and T Stürmer. "Indications for propensity scores and review of their use in pharmacoepidemiology." Basic & clinical pharmacology & toxicology 98.3 (2006): 253-259.

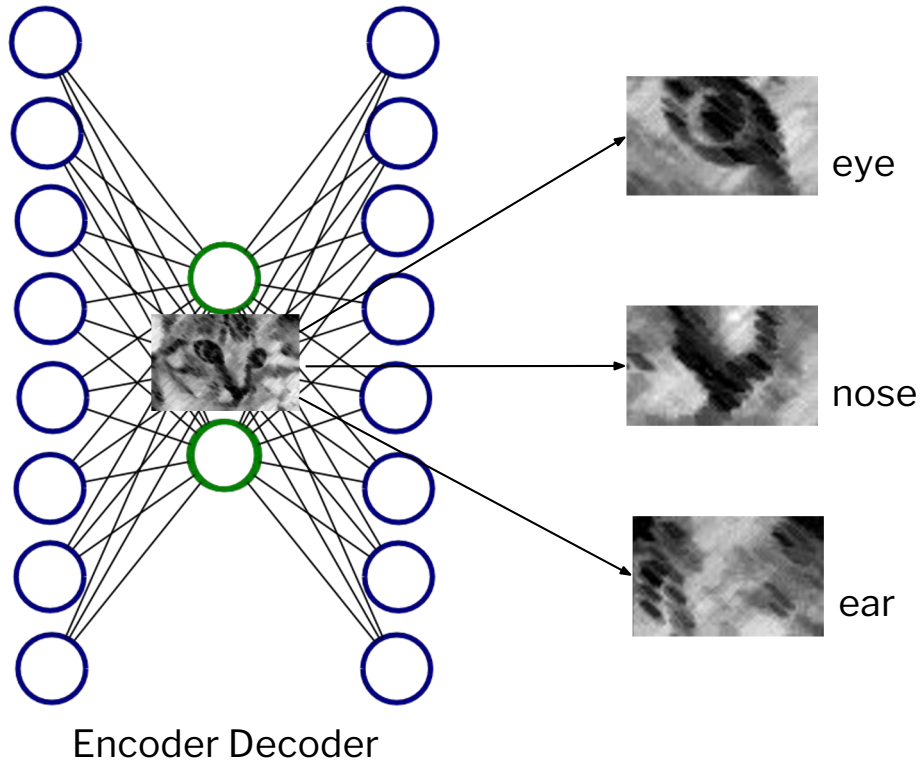
Autoencoders

Unsupervised (*self-supervised*) deep learning architecture

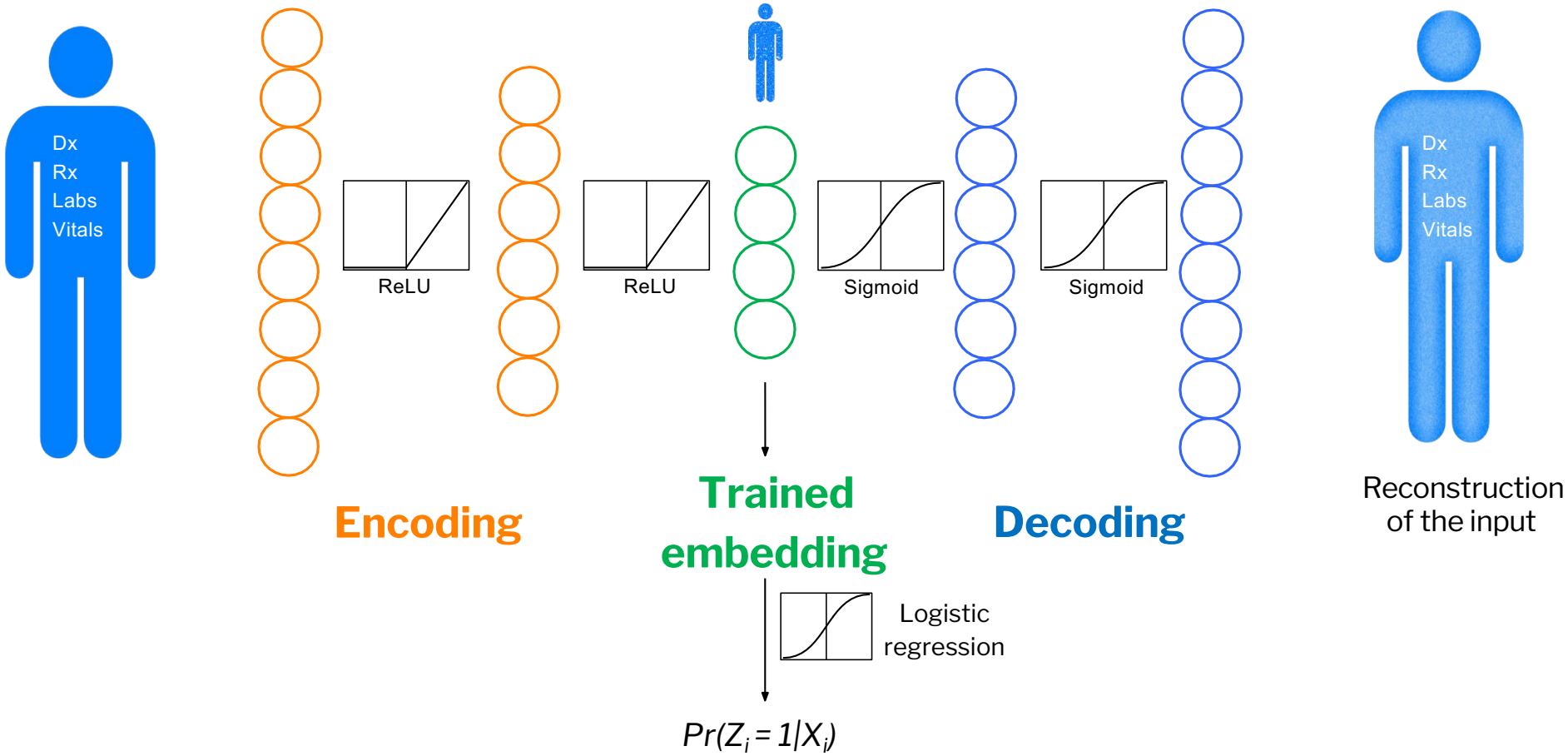


Autoencoders

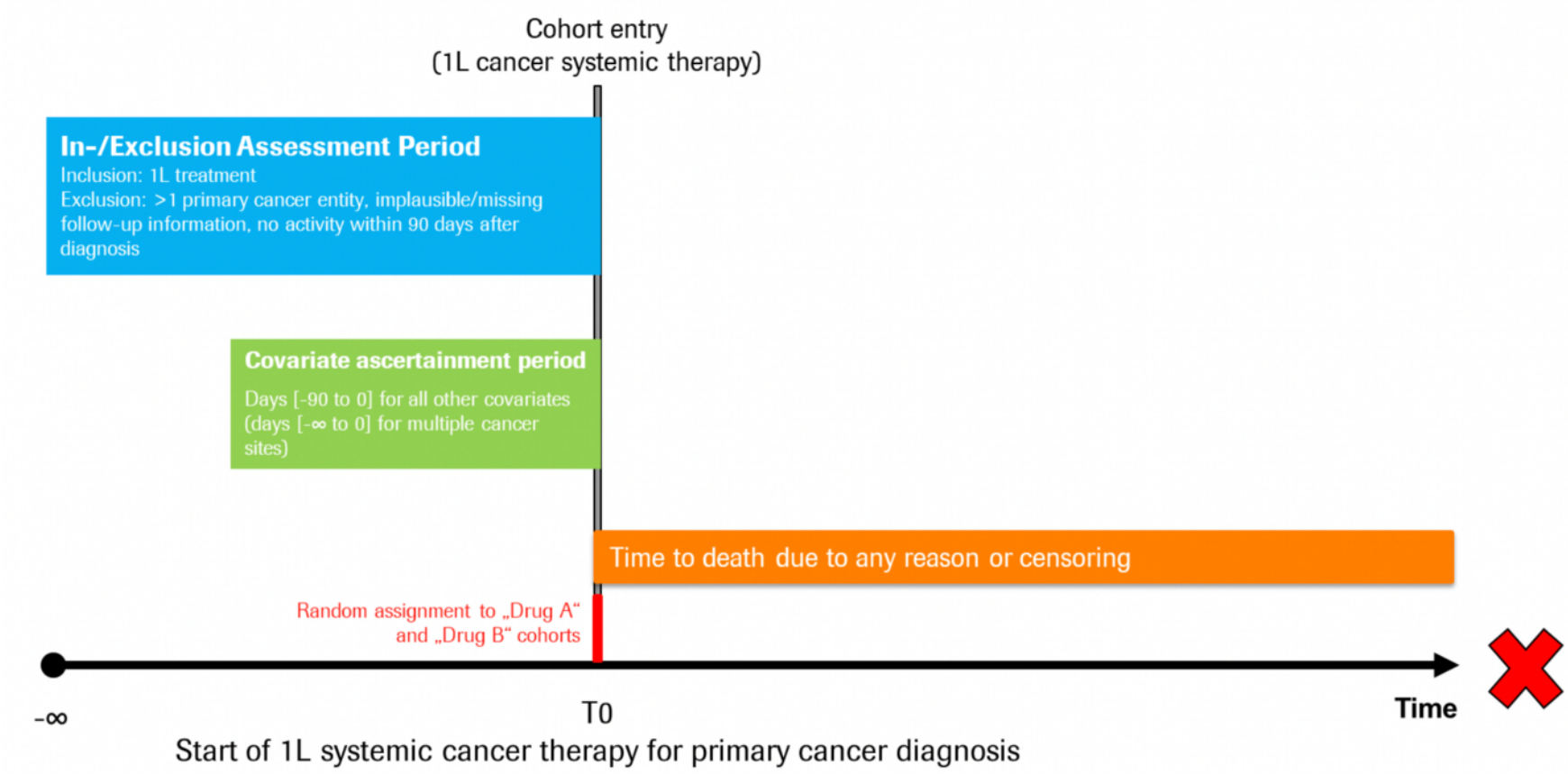
Unsupervised (*self-supervised*) deep learning architecture



Learning latent patient representations using unsupervised autoencoders

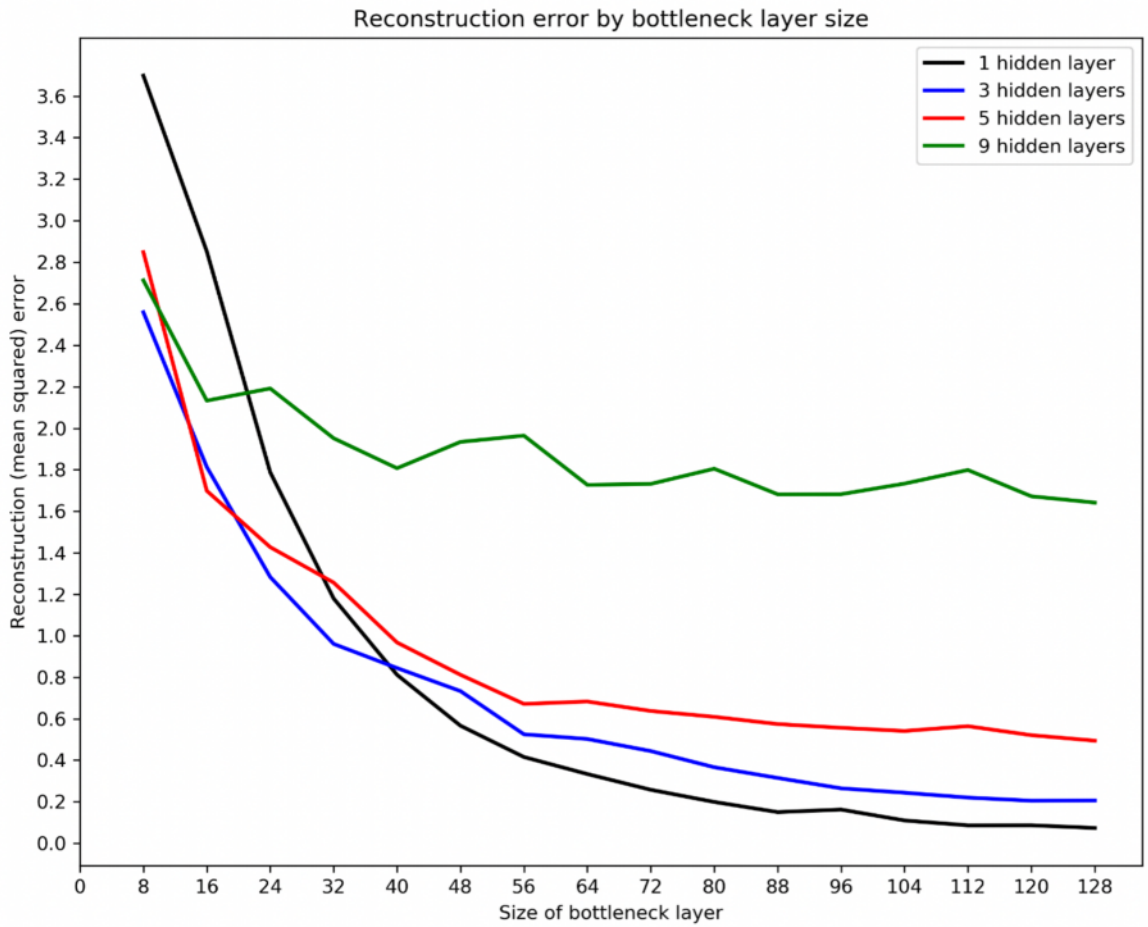


Study design (adapted from Schneeweiss S. et al. Ann Intern Med. 2019 Mar 19;170(6):398-406)



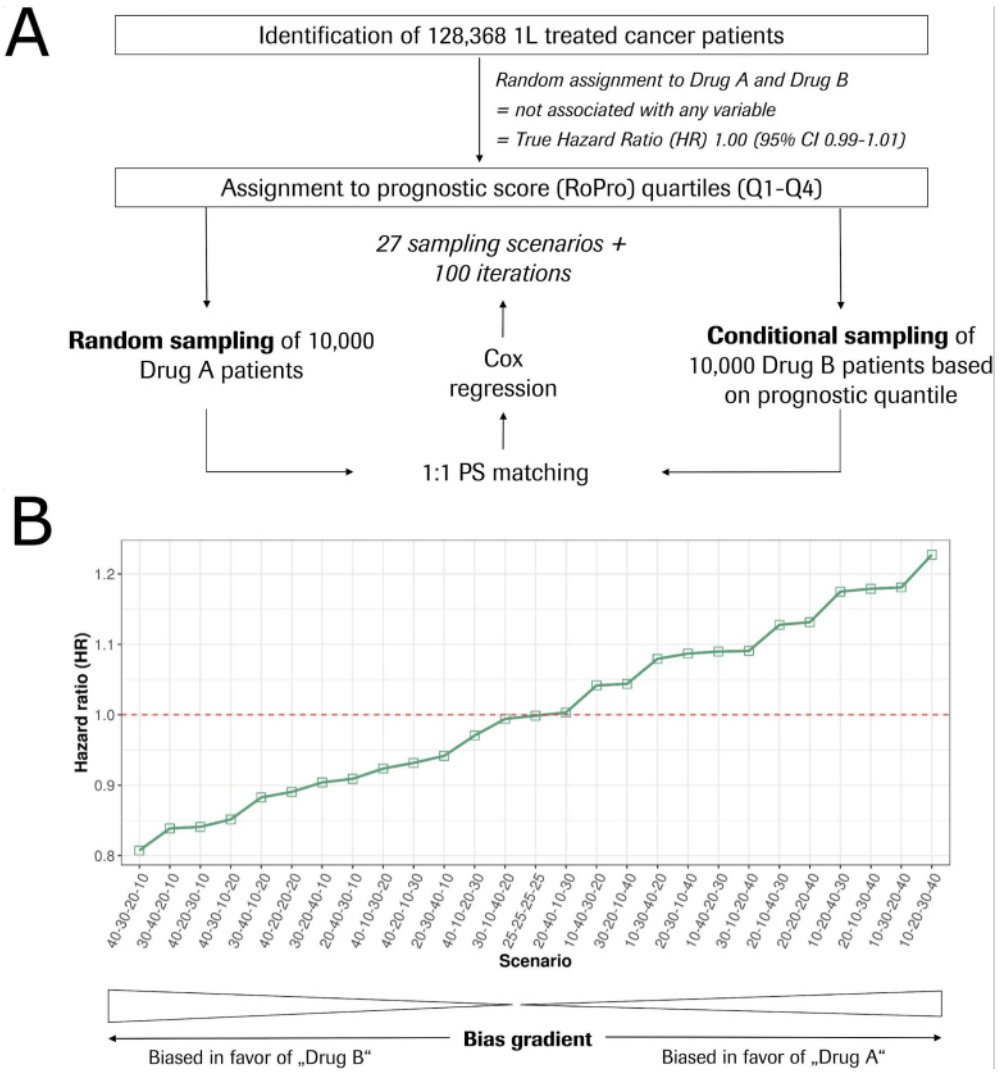
Training and hyperparameters

eFigure 10. Autoencoder development – reconstruction error by number of hidden layers and bottleneck layer size (318 total covariates).



Hyperparameters:

- Adadelta optimizer with adaptive learning rates
- Activation functions:
  - Encoding: Rectified linear units (ReLu's)
  - Decoding: Sigmoid
- Binary cross-entropy loss function
- 3% Noise
- Model checkpoint callbacks
- Batch size 64 (# samples before updates to model weights)
- 256 epochs (# total passes per batch size)
- # of hidden layers and bottleneck size was determined through grid search by minimizing reconstruction error (MSE) between original and predicted values



- Correlation between prognostic score–based balance measures for propensity score models with bias in the treatment effect estimate[1]
- Use of ROPRO score[2] to induce some “artificial imbalance” based on the conditional sampling of patients with different risk quartiles in the control group (Drug B cohort)

- 27 different samplings scenarios considered
- Different magnitudes of imbalance in both directions
- Objective: Performance of different propensity score models to adjust for imbalances and recover true HR = 1.00
- Performance metrics: SMD, RMSE, %bias, CI coverage

[1]*Stuart EA, Lee BK, Leacy FP. J Clin Epidemiol. 2013;66(8 Suppl):S84-S90.e1*
[2]*Becker T, Weberpals J, Jegg AM, et al. Ann Oncol. 2020*

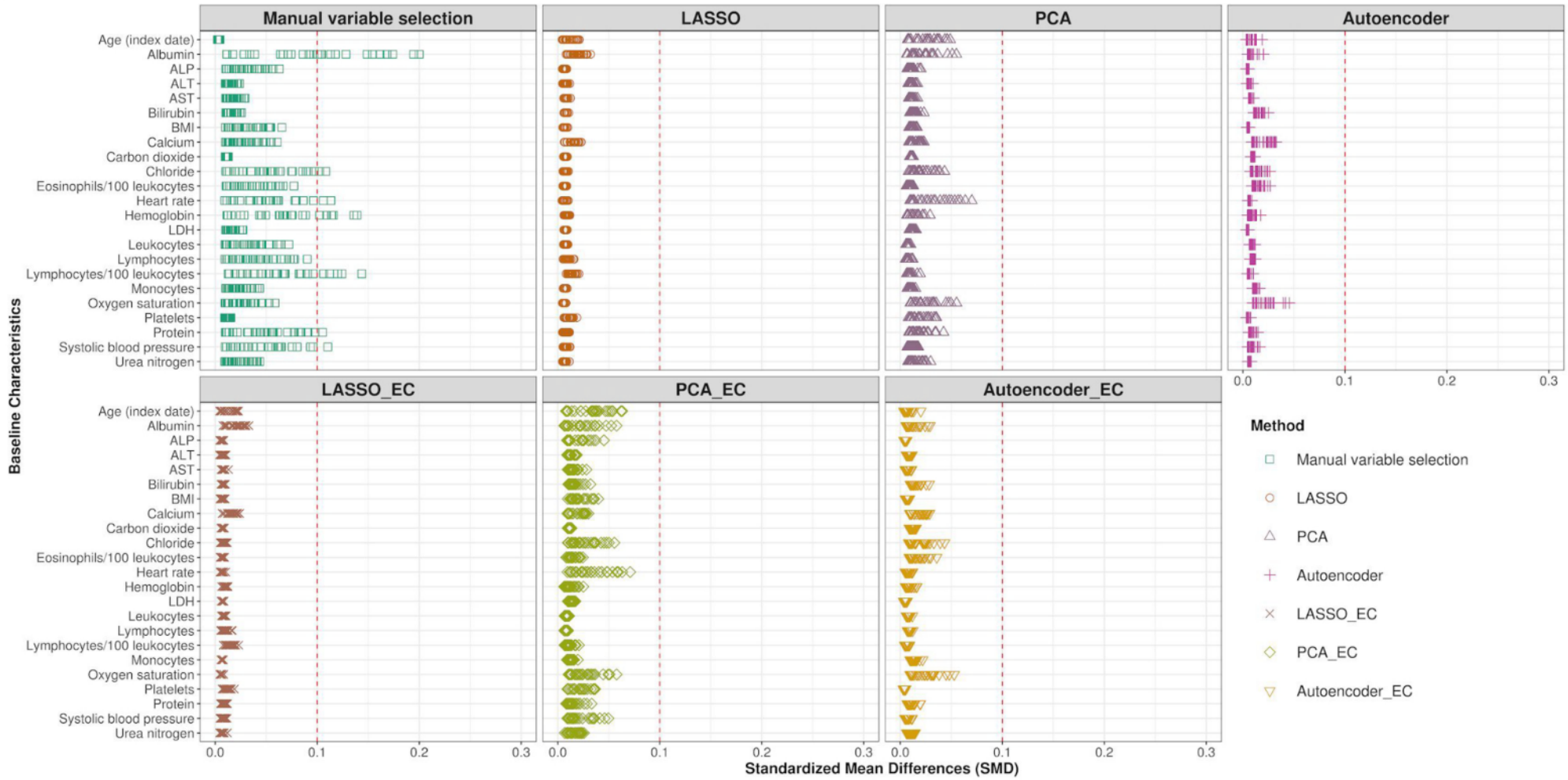
Table. Models and adjustment strategies compared in simulation framework.

Model	Adjustment strategy[a]	Data-adaptive covariate selection / transformation	Covariates adjusted for or potential covariates to choose from
1	Unadjusted	-	-
2	Multivariable Regression (direct outcome model)	No	Age, cancer entity, gender, stage, histology, healthcare provider, race/ethnicity, time from initial cancer diagnosis to 1L initiation, calendar year of initial cancer diagnosis
3	Manual variable selection	No	Age, cancer entity, gender, stage, histology, healthcare provider, race/ethnicity, time from initial cancer diagnosis to 1L initiation, calendar year of initial cancer diagnosis
4	LASSO	Selection	All generally available covariates. Algorithm picks covariates according to shrinkage/regularization
5	PCA	Transformation	All generally available covariates. Algorithm computes linear transformation of all covariates in a dataset to principal components (PCs) of which the top *n* PCs, explaining 80% variance, were chosen
6	Autoencoder	Transformation	All generally available covariates. Algorithm computes lower-dimensional representation of j dimensions based on non-linear data operations into latent-space variables
7	LASSO EC	Transformation	Model 4 + 123 empirical covariates[c]
8	PCA EC	Selection	Model 5 + 123 empirical covariates[c]
9	Autoencoder EC	Transformation	Model 6 + 123 empirical covariates[c]

Abbreviations: 1L = first-line systemic cancer treatment, EC = Empirical covariates, LASSO = Least absolute shrinkage and selection operator, PC(A) = Principal component (analysis)

[a] In model 2 the estimate is directly computed from a multivariable regression while models 3-9 are based on propensity score matching
[b] Total of 318 demographic, clinical, cancer-/disease-specific covariates
[c] Total of 123 frequency covariates derived from step 1-3 of the hdPS algorithm

Simulation results - balancing

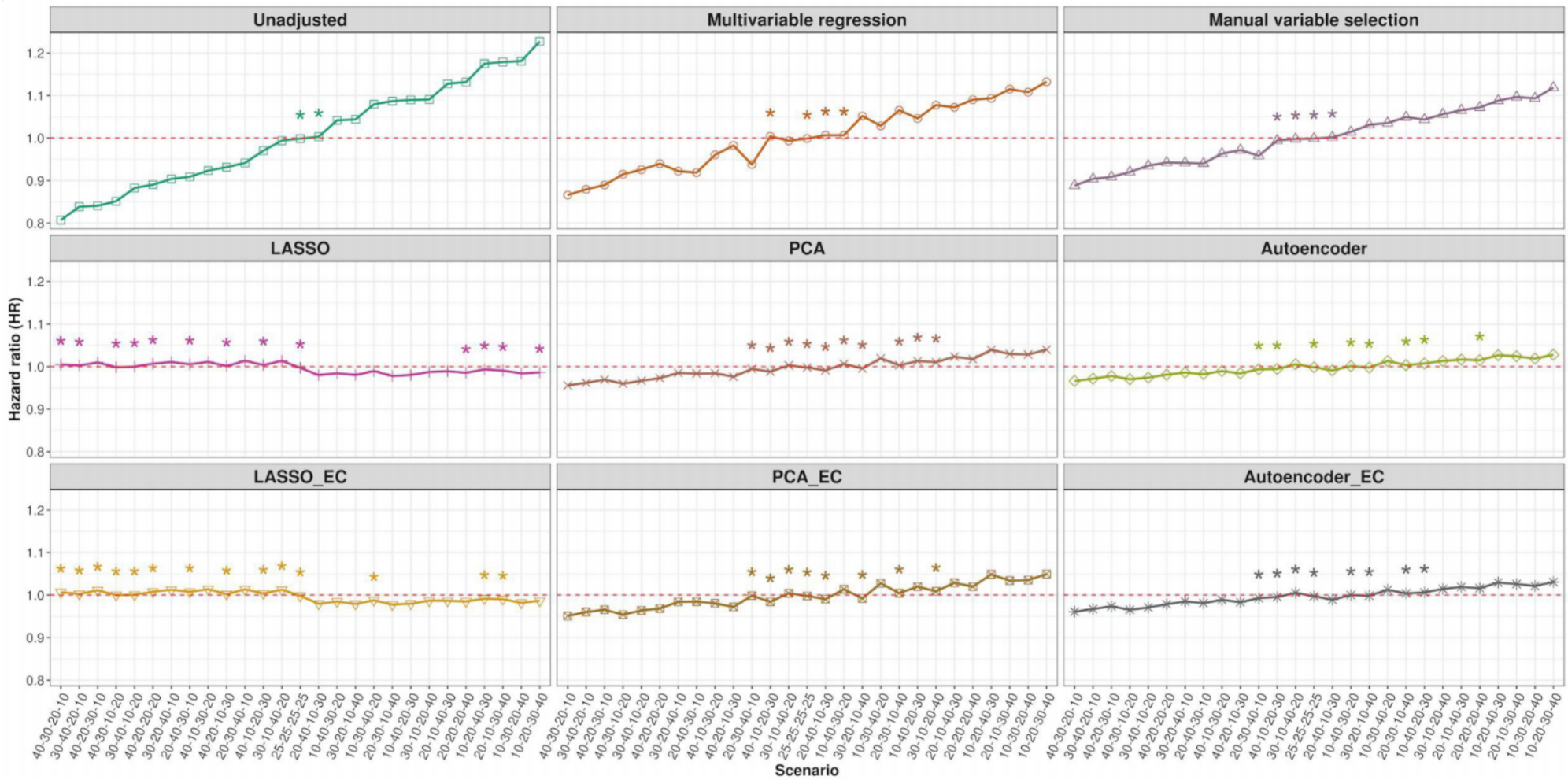


Simulation results - confounding adjustment

Table. Summary of adjustment performance across all scenarios.

Method	RMSE	Bias (%)	CI coverage (%)
1. Unadjusted	0.1205	10.4	16.41
2. Multivariable regression	0.0790	6.75	27.67
3. Manual variable selection	0.0670	5.73	32.81
4. LASSO	0.0205	1.65	93.74
5. PCA	0.0293	2.39	79.59
6. Autoencoder	0.0248	2.00	87.70
7. LASSO EC	0.0210	1.69	93.52
8. PCA EC	0.0329	2.71	74.00
9. Autoencoder EC	0.0265	2.15	85.19

Abbreviations: CI = Confidence interval, EC = Empirical covariates, LASSO = Least absolute shrinkage and selection operator, PC(A) = Principal component (analysis), RMSE = Root mean squared error



------ *Dashed line represents true HR*

Case study to illustrate application of autoencoder-derived propensity score:

Emulation of PRONOUNCE target trial

Table. Summary and comparison of main protocol elements between PRONOUNCE RCT* and emulated target trial using autoencoder-derived propensity score

***PRONOUNCE RCT[1] (2015)**

- Randomized, open-label, phase III trial
- Non-small cell lung cancer
- Setting: 1L
- Intervention: carboplatin/pemetrexed followed by pemetrexed maintenance versus bevacizumab/ carboplatin/paclitaxel followed by bevacizumab maintenance
- $HR_{PRONOUNCE}$ 1.07 (95% 0.83-1.36)

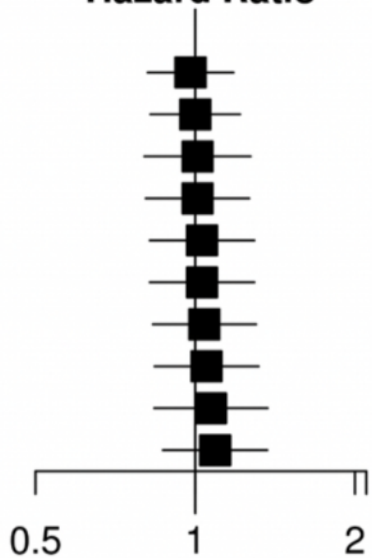
Protocol component	PRONOUNCE	Target trial emulation
Eligibility criteria[1]	• Chemotherapy naïve/1L • ≥18 years of age • NSCLC • Stage IV • histologically or cytologically confirmed nonsquamous • ECOG PS 0 or 1 • measurable disease by Response Evaluation Criteria in Solid Tumors and adequate organ function were eligible • Actual primary completion date: January 2013 • Results first posted on clinicaltrial.gov: April 2014 • Referenced paper published: January 2015	• 1L (exclusion of patients with no activity 90 days after diagnosis to mitigate chances for 1L misclassification) • ≥18 years of age • Diagnosed with lung cancer (ICD-9 162.x or ICD-10 C34x or C39.9) with at least two documented clinical visits on or after January 1, 2011 • + Diagnosed with Stage IIIB, IIIC, IVA or IVB NSCLC on or after 1/1/2011, or diagnosed with early-stage NSCLC and subsequently developed recurrent or progressive disease on or after 1/1/2011 • Non-squamous histology • Treatment initiation before October 2016[2] • No EGFR and ALK genomic aberration[3]
Treatment strategies	Carboplatin/pemetrexed followed by pemetrexed maintenance vs. bevacizumab/carboplatin/paclitaxel	Carboplatin/pemetrexed followed by pemetrexed maintenance vs. bevacizumab/carboplatin/paclitaxel
Assignment procedures	Random assignment to either treatment strategy in a 1:1 ratio	Propensity score matching in a 1:1 ratio (nearest neighbor without replacement)
Follow-up period	Date of randomization to the date of death from any cause or censoring at the last date the participant was known to be alive.	Date of initiation of the respective 1L maintenance therapy (= first possible time to meet all inclusion criteria) to date of death from any cause or censoring at the last confirmed structured activity
Outcome	Overall survival (secondary outcome in original trial)	Overall survival
Causal contrasts of interest	Intent-to-treat effect	Counterfactual comparison of initiators of the two different treatment strategies (observational equivalent of the intent-to-treat analysis)

Abbreviations: 1L = first-line systemic cancer treatment
[1] Only major eligibility criteria for PRONOUNCE are displayed
[2] In October 2016 the first checkpoint inhibitor for 1L NSCLC was approved
[3] The target trial population is restricted to EGFR and ALK negative patients as NSCLC patients with EGFR and ALK aberrations usually experience different treatment strategies.

[1] Zinner RG, Obasaju CK, Spigel DR, et al. PRONOUNCE: randomized, open-label, phase III study of first-line pemetrexed + carboplatin followed by maintenance pemetrexed versus paclitaxel + carboplatin + bevacizumab followed by maintenance bevacizumab in patients with advanced nonsquamous non-small-cell lung cancer. J Thorac Oncol. 2015;10:134–142.

Case study results

PS estimation method	N	Events	HR	95% CI
Unadjusted	781	606	0.98	0.82-1.19
Multivariable regression	781	606	1.00	0.82-1.21
Autoencoder	372	291	1.01	0.80-1.27
LASSO	372	297	1.01	0.81-1.27
PCA EC	372	293	1.03	0.82-1.29
LASSO EC	372	300	1.03	0.82-1.29
PCA	372	297	1.04	0.83-1.30
Manual variable selection	372	296	1.05	0.84-1.32
PRONOUNCE trial	361		1.07	0.83-1.36
Autoencoder EC	372	297	1.09	0.87-1.37



Favors Carbo,Pem *Favors Beva,Carbo,Pac*

FIGURE 7. Forest plot illustrating HRs and 95% confidence intervals (CIs) for overall survival by PS estimation method. HR indicates hazard ratio; LASSO, least absolute shrinkage and selection operator; PCA, principal component analysis.

Conclusions & Outlook

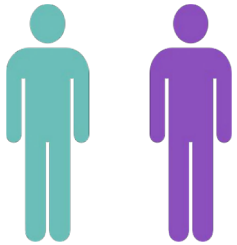
- For both prediction and inference models, deep learning worked well but not substantially better than established methods
- Given time and resources, one should consider if it's worthwhile tuning neural networks for tabular data versus using tree-based or penalized regression models
- Situation may be different for tabular time-series data or when it comes to enrich tabular data with more complex and less sparse data (e.g. images, single-cell seq, unstructured [notes], etc.)
- Outlook:
    - Test autoencoder algorithm in multimodal databases for data enrichment
    - Optimize DL loss functions to target causal inference questions (e.g. optimize towards cohort balancing, doubly robust models, etc.)

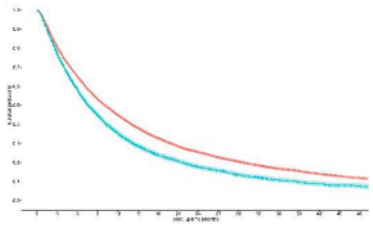
Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay





Resources and code availability

Papers can be found at:

https://www.ncbi.nlm.nih.gov/sites/myncbi/1heTeqnwOBzQ5/collections/62116377/public/

Deep learning prognostic scores (including ROPRO)

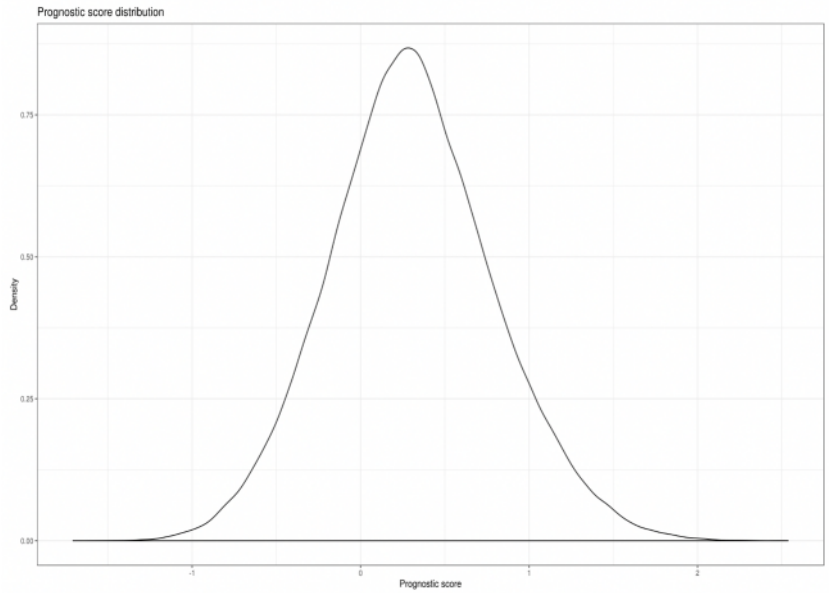
- Analysis code and files are published in supplement of manuscript at https://www.frontiersin.org/articles/10.3389/frai.2021.625573/full

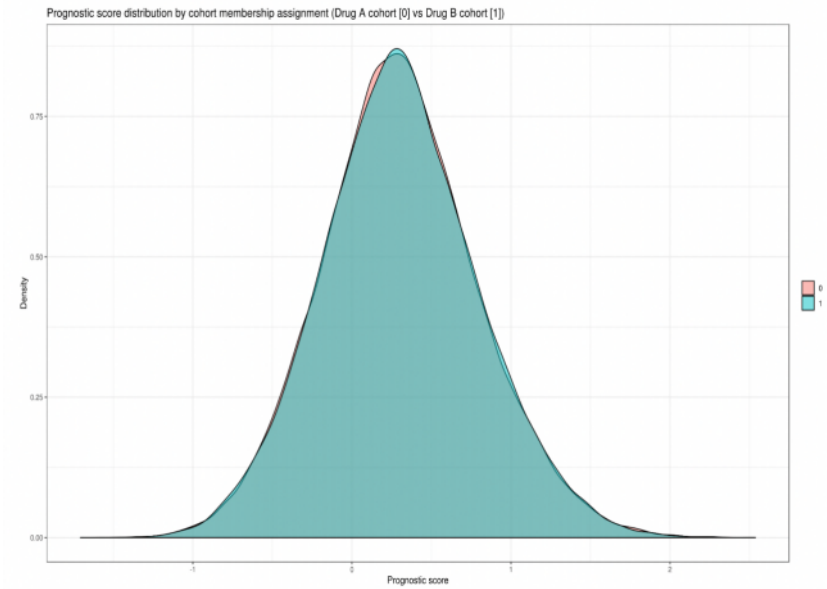
Deep learning propensity scores scores

- Code for autoencoder training and simulation is published at https://github.com/janickweberpals/autoencoderPS

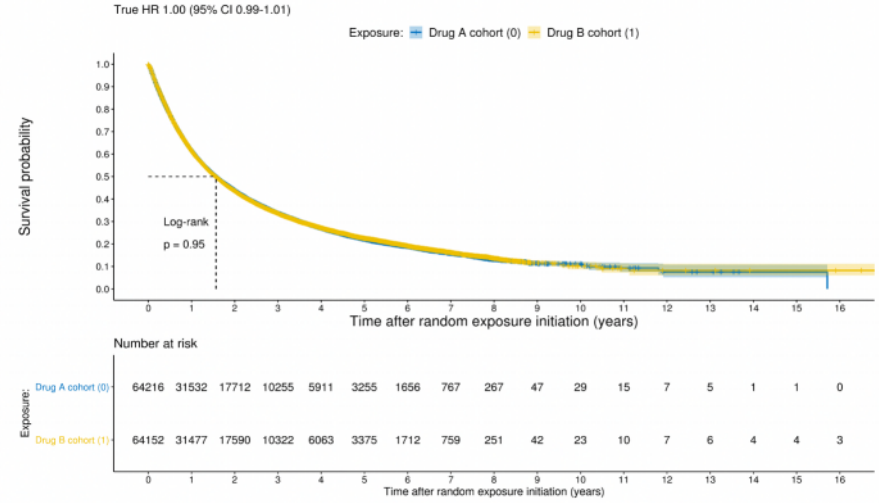
jweberpals@bwh.harvard.edu

janickweberpals.github.io





eFigure 2. Baseline survival estimates in entire study population by cohort assignment to Drug A vs Drug B.



eFigure 3. Kaplan-Meier survival estimates stratified by prognostic score quartile.