Title: Attention based deep neural networks in patient level prediction

PRESENTER: **Egill Fridgeirsson**

INTRO:

- Recently there have been rapid advances using attention-based models in deep learning[1]
- In attention the model learns relations between representations of the input features
- Here we test whether attention-based models can outperform strong linear and non-linear baselines on a diverse set of tasks

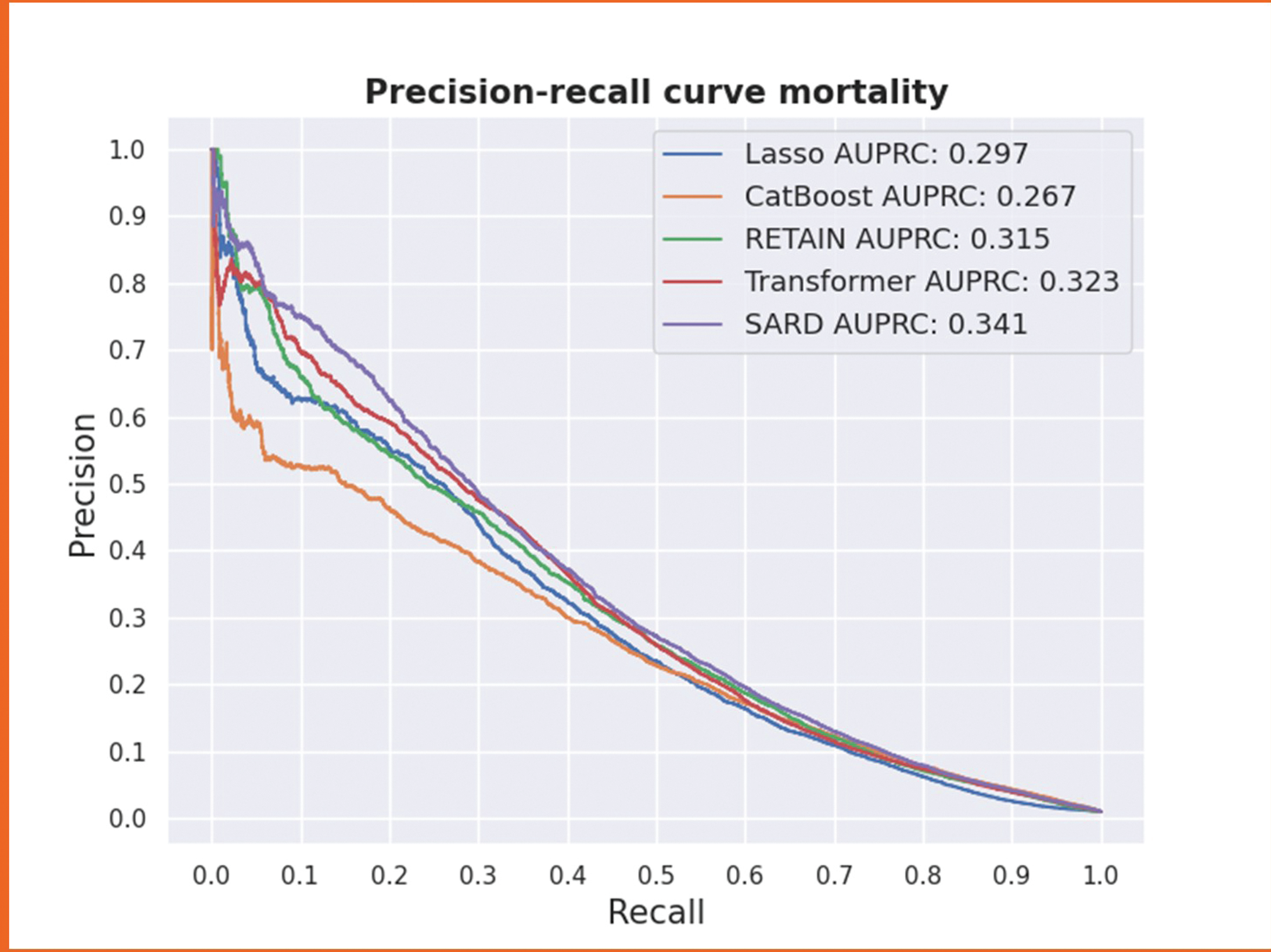
METHODS

1. We test two models, RETAIN[2] which is a recurrent neural network with attention on the hidden states.
2. We also test a transformer which is a pure attention-based model
3. We test a transformer both from scratch and using reverse distillation (SARD) where it learns from a strong linear baseline model[3].
4. The two baselines are an L1 regularized linear model (LASSO) and gradient boosted trees (catboost)
5. We test on three tasks on data from the IPCI (www.ipci.nl) database from the Netherlands:
    1. Mortality within 30 days from GP visits of patients older than 60.
    2. Dementia in next 5 years after a GP visit in 2012-2014 of patients aged between 50-79
    3. Readmission within 30 days after an inpatient visit of adults.
6. Conditions, procedures and drug exposure are extracted from the year before the index visit.
7. We use the PatientLevelPrediction[4] (PLP) package to extract features, we remove features occurring in less than 0.1% of patients/visits and normalize continuous features.
8. We use a 50-25-25 split for training-validation-test sets
    1. For Lasso we use a grid search with variances from 0.01-20.
    2. For all other models we use a randomized search with 100 iterations to select best hyperparameters on validation set

AUC (95% CI)	Mortality	Readmission	Dementia
LASSO	0.902 (0.001)	0.636 (0.07)	**0.869 (0.1)**
Catboost	0.931 (0.003)	0.635 (0.007)	0.865 (0.01)
RETAIN	0.923 (0.003)	0.632 (0.07)	0.857 (0.02)
Transformer	0.926 (0.003)	0.643 (0.007)	0.860 (0.01)
SARD	0.931 (0.003)	**0.644 (0.007)**	**0.869 (0.01)**

AUPRC	Mortality	Readmission	Dementia
LASSO	0,297	0,176	**0,088**
Catboost	0,267	0,175	0,082
RETAIN	0,315	0,166	0,075
Transformer	0,323	0,179	0,08
SARD	**0,341**	**0,183**	0,084

- Overall the performance is similar (< 1%) with regards to the AUC
    - Except LASSO is worse in mortality prediction
- The deep learning models are competitive to the baselines and SARD is either equal or slightly better than the baselines in terms of AUC.
- Reverse distillation improves the model over training from scratch
- With regards to the AUPRC which better reflects performance for the outcome (minority) class SARD is better than others in mortality prediction
- Overall the baselines are competitive but there seems to be slight improvements in precision recall with SARD



Refs

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv. Neural Inf. Process. Syst., vol. 2017- Decem, 2017,p. 5999–6009. https://doi.org/10.5555/3295222.3295349
2. Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. Adv Neural Inf Process Syst 2016:3512–20
3. Kodialam RS, Boiarsky R, Lim J, Dixit N, Sai A, Sontag D. Deep Contextual Clinical Prediction with Reverse Distillation. Proc AAAI Conf Artif Intell 2020;35:249–58.
4. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Informatics Assoc 2018;25:969–75. https://doi.org/10.1093/jamia/ocy032

Data information

	Mortality	Readmission	Dementia
Target cohort	3.802.717 visits	220.580 visits	169.595 patients
Outcome (%)	36.922 (1%)	25.163 (11.4%)	2370 (1.4%)
Index event	GP visit after 60	Inpatient visit of adults	GP visit in 2012-2014 of patients aged 50-79
Time-at-risk	30 days	30 days	5 years
Observation window	1 year prior to index	1 year prior to index	1 year prior to index

- We use the same train-test splits from the PLP package for all models
- Non temporal features are concatenated to visit embeddings for the deep models
- The transformer uses sinusoidal position embeddings

Code available at:
https://github.com/mi-erasmusmc/Sard

Work will eventually be part of the deepPLP package at:
https://github.com/OHDSI/DeepPatientLevelPrediction/

AUTHORS: Egill Fridgeirsson, David Sontag, Peter Rijnbeek

Erasmus MC University Medical Center Rotterdam

OHDSI