From baseline data to outcomes: are artificial intelligence based models really competitive?

Lidia Sacchetto[1,2], Mauro Gasparini[1], Karl Köchert[3]

[1]Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy.
[2]Department of Mathematics, Universitá degli Studi di Torino, Turin, Italy.
[3]Department of Clinical Statistics EU, SBU Oncology, Pharmaceuticals, Bayer AG, Berlin, Germany.

BACKGROUND

- In clinical trials a huge amount of clinical information is routinely collected for each subject, with a large investment of time and resources.
- Only a small fraction of these data is commonly used in standard analyses, for stratification and regulatory assessment purposes.
- The popularity of Big Data and Artificial Intelligence (AI) approaches in different fields paves the way to adopt and adapt these methods for extracting useful information from clinical data.

AIM

To develop multivariate predictive models for treatment efficacy with time-to-event outcomes (overall survival), using only baseline data routinely collected in clinical trials and applying different methods to understand if there is valuable information not accounted for by standard methodology.

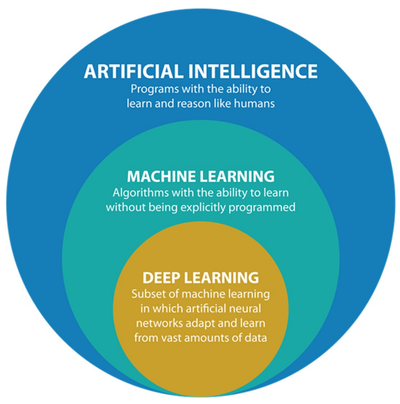
MATERIALS & METHODS

Data from a randomized, double blind, placebo-controlled, multicenter phase III study.

- Dataset1: 573 subjects, 184 clinical variables (including only baseline measurements);
- Dataset2: 499 subjects, 450 variables (about 260 variables on proteins information).

Comparison of different methodologies (using the coefficient of determination R^2, i.e. the percentage of variance explained by the model)

- Traditional low-dimensional Cox model on a selection of 8 variables medically relevant for the disease.
- Machine learning approaches (survival and regression Random Forests (RFs)) on all collected variables.
- Deep learning algorithms, as shallow feed-forward Neural Networks (NNs) (also including an additional outcome) on all collected variables.



RFs and NNs involve a number of hyperparameters to be tuned.

- Grid search approach to find the optimal combination which provides the best performance measure.
- Test sets, out-of-bag measures and cross-validation techniques adopted to reduce overfitting in results.

Data pre-processing

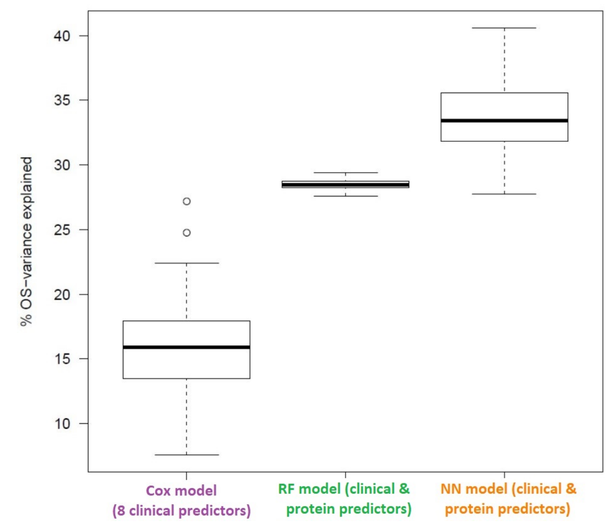
- Min-Max transformation to rescale all the continuous variables in $[0, 1]$.
- Missing imputation on variables with less than 25% of missing values, using random forests; exclusion of variables with higher percentages of missing values.
- One-Hot-Encoding transformation for categorical variables (with more than two levels).

Softwares

- Cox model and Random Forests implemented in R (packages: survival, ranger, randomForestSRC, missForest).
- Neural Networks built in Python (libraries: Keras, TensorFlow, scikit-learn).

RESULTS

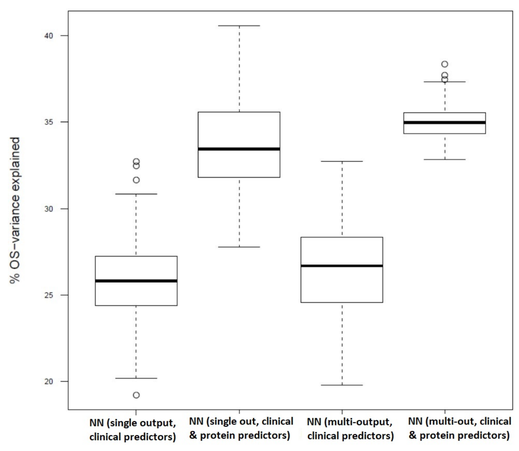
- Substantial improvement in the percentage of variance explained by the models, especially after the inclusion of proteins information*



Method	$R^2(\%)$	$[Q1 - Q3]$
Cox model	15.9	13.5 – 17.9
Regression RFs	28.5	28.3 – 28.7
Neural Networks	33.4	31.8 – 35.5

Single *versus* multi-output networks*

- The simultaneous analysis of an additional output (minimum percentage change in tumor volume from the baseline measurement) allows explaining an additional 2% of the overall survival related variance.



Boxplots obtained repeating estimations 100 times.

DISCUSSION

Strengths of the study

- First attempt to look in depth into all clinical data routinely collected in a trial to understand if there is valuable unused information.
- Different approaches adopted: from traditional to AI models.
- Models easily applicable to other databases (given in the appropriate format).

Limitations of the study

- Only one clinical database analysed.
- Hyperparameters tuning to be refined for the multi-outputs neural network framework.
- Possible residual unknown amount of overfitting: overoptimistic results and lack of reproducibility.

Future steps

- To predict the clinical outcome of interest for each subject (personalized information) using developed models and to provide variables importance measures to give indications on their usefulness.

CONCLUSIONS

- **Worthy information related to subject clinical outcome is already contained in baseline measurements.**
- **Random forests and neural networks allow to better explain OS related variance (using all the available variables).**
- **The advantage of AI-based models in term of performance and exploitation of data compensates higher complexity and longer time needed for predictions.**