Model Ensembling vs Data Pooling: Alternative ways to merge hospital information across sites

Kristin Corey[1,2], Elizabeth Lorenzi[2,3], Michael Gao[1], Suresh Balu M.B.A[1, 2], Mark Sendak M.D. M.P.P.[2]
[1]Duke University School of Medicine, [2]Duke Institute for Health Innovation, [3]Department of Statistical Science, Duke University Health System

Background. Healthcare data from institutions across distributed research networks are increasingly utilized as sources for research due to the nationally representative population in the data. Pragmatic trials are conducted in this setting and there is a valuable opportunity for machine learning models to be built within distributed research networks. However, joining a network to either centralize or federate data is expensive and time consuming due to the complexities of sharing data and insights derived from data. In response to this problem, we conducted experiments utilizing data pooling and ensemble learning by data pooling across three hospital sites (Hospitals A, B, C) serving heterogenous populations within a single healthcare system. The goal of the analysis is to characterize how model performance can change for sites that join a distributed research network. In this abstract, we highlight one set of our experiments using one target hospital.

Methods. Using models developed to predict post-operative complications[1], we conducted three experiments, each comparing LASSO and extreme gradient boosted decision tree models. All experiments used a stable testing cohort of 1500 encounters from hospital site C that were excluded from the training of the models, and all models were cross-validated to find the optimal hyperparameters. The first experiment analyzed both models trained and tested on the same site. The second experiment analyzed both models trained on all pooled data across three sites and tested on one of the sites. Lastly the third experiment analyzed ensembling models trained from each of the three sites without data pooling and tested on one of the sites, where the predictions from each site-specific model are averaged.

Results. Table 1: Experiments tested using LASSO model

Training Cohort	Testing Cohort	Model Type	AUROC	AUPROC
C	C	gradient boosted decision trees	0.779 (0.742, 0.815)	0.378
A, B, & C	C	gradient boosted decision trees	0.802 (0.766, 0.837)	0.374
Ensembled Models from A, B, and C	C	gradient boosted decision trees (3 models)	**0.858 (0.827, 0.888)**	**0.425**

Table 2: Experiments tested using extreme gradient boosted decision trees

Training Cohort	Testing Cohort	Model Type	AUROC	AUPROC
C	C	LASSO	0.766 (0.729, 0.803)	0.360
A, B, & C	C	LASSO	**0.806 (0.773, 0.840)**	0.345
Ensembled Models from A, B, and C	C	LASSO (3 models)	**0.803 (0.766, 0.838)**	**0.412**

Conclusion. These experiments demonstrate that pooling data can increase model performance for both LASSO and extreme gradient boosted decision trees. However, ensemble methods demonstrate equal model performance for LASSO and superior performance for extreme gradient boosted decision trees. This is most likely due to the fact that ensembling allows for the models to explicitly learn from each individual hospital, whereas data pooling treats all observations as stemming from the same site. Extreme gradient boosted decision trees are able to learn complex relationships between covariates and the outcome resulting in better performance compared to LASSO, especially when used in ensemble methods. These experiments and results have large implications for health care delivery systems and distributed research networks. If ensembling models is a feasible way to leverage information across network sites, distributed research networks that require data pooling may face significant challenges. Individual sites may choose to build local models that distributed research networks can then ensemble and scale across networks. Further research is needed to evaluate the value of data pooling and ensemble methods for additional disease conditions across additional sites.

[1]Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. (2018) development and validation of machine learning model to identify high-risk surgical patient using automatically curated electronic health record data (Pythia: A retrospective, single-site study. PLoS Med 15(11): e1002701.