Machine Learning, Individualized Modeling, and Risk Factor Analysis of Clinical Diagnoses

Xinyi Zhang[1]; Tara Hashemian[2]; Jason Cory Brunson[3]

[1]College of Liberal Arts and Sciences, [2]Department of Biostatistics, [3]Laboratory for Systems Medicine

INTRODUCTION

Motivation

Conventional prediction tools have complementary strengths and weaknesses. We propose to couple these techniques in order to achieve improved accuracy while preserving interpretability. Models fit "locally" to similarity-based cohorts, rather than "globally" to whole data sets, have shown improvements in prediction accuracy and also have the potential for individualized interpretation, e.g. identifying key risk factors and estimating their effects.

	Regression	Machine Learning	Case-Based Reasoning
Interpretability	Moderate	Low	High
Accuracy	Moderate	High	Low

Objectives

1. Compare the performance of global and individualized models on several clinical diagnosis problems.
2. Compare individualized importance measures of clinical risk factors between subpopulations.
3. Identify patient phenotypes based on risk factor importance.

MATERIALS & METHODS (PART I)

Data

We used 3 data sets from the University of California Irvine Machine Learning Repository [1].

Clinical Domain	Task	Number of Observations	Number of Attributes
Dermatology	Classify dermatological presentations to one of 6 erythemato-squamous diseases	366	34
Breast Cancer	Classify breast tumors as benign or malignant	569	11
Heart Disease	Identify the presence of heart disease	270	13

Similarity Measure

A patient similarity measure is a numerical value calculated from case-level elements in biomedical data. It is used to quantify the relevance of one patient or case to another.

We considered two choices of similarity measure:

- Euclidean Distance: $d(p,q) = \| p - q \|$
- Cosine Similarity: $s(x,y) = (x \cdot y)/(\| x \| \cdot \| y \|)$

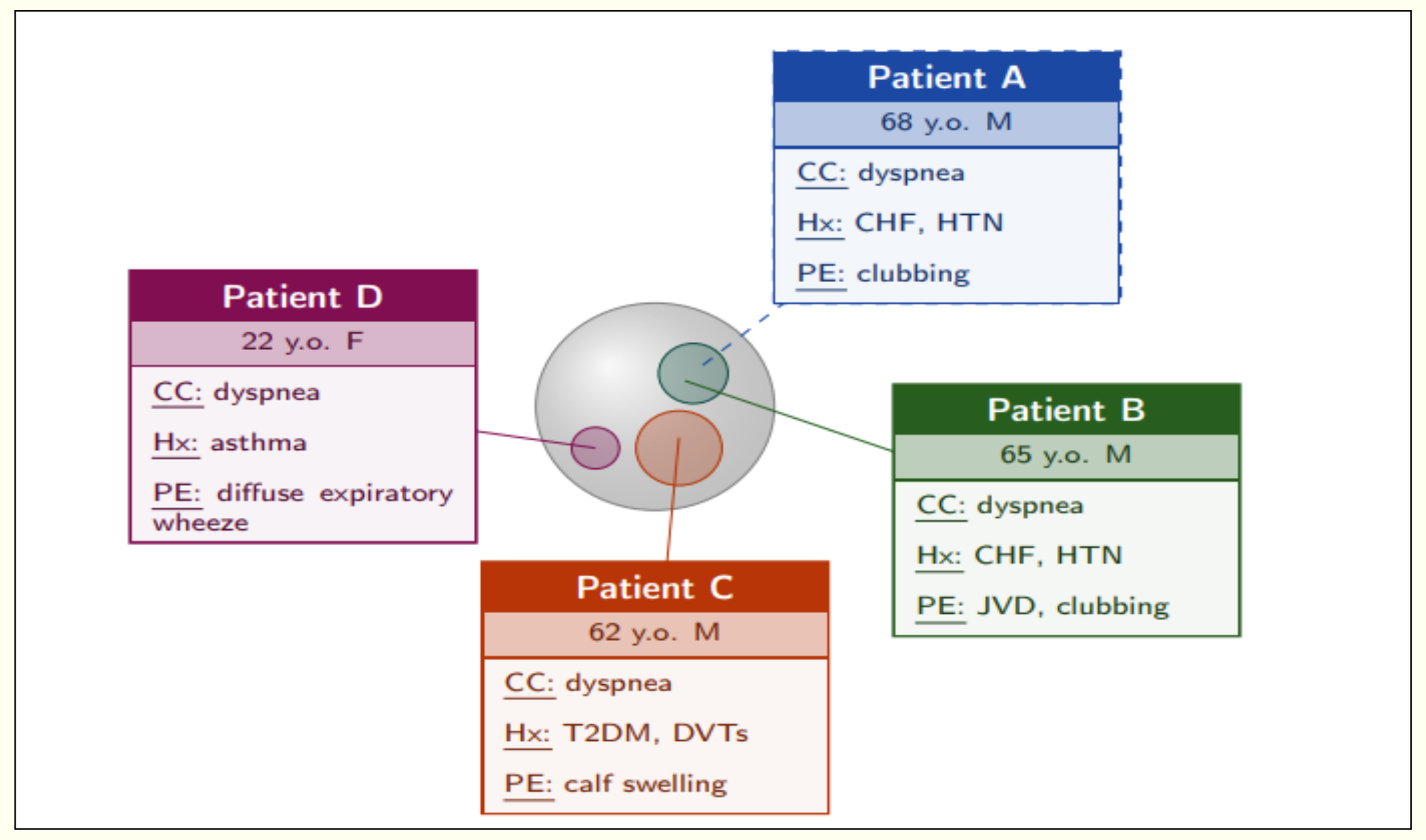


Figure 1. Patient similarity measurement and similarity cohorts

METHODS (PART II)

Individualized Model

1. New cases present
2. For every case, extract clinical features
3. Compute the similarity between new cases and cases in a corpus
4. Retrieve a cohort of most similar cases for each new case
5. Fit the statistical model to the similarity cohort
6. Generate a prediction for each new case

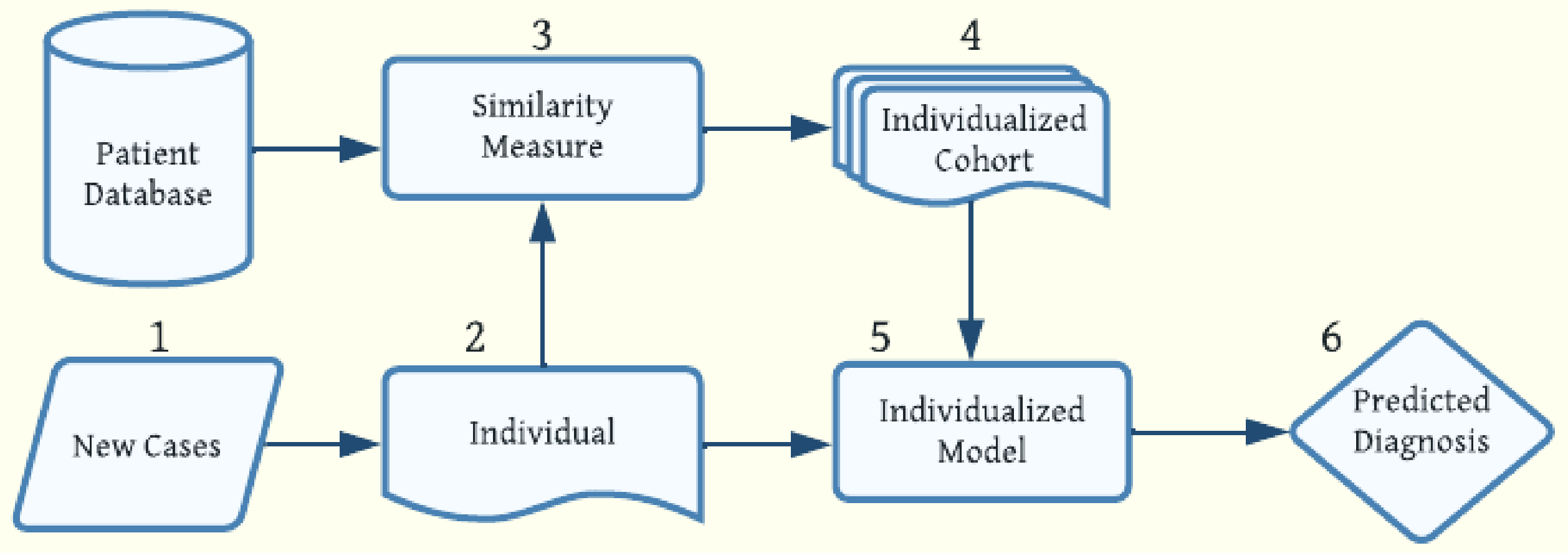


Figure 2. Individualized modeling procedure

Workflows

- We built ML workflows in R using the tidymodels collection [2]
- Conventional families of prediction models included logistic regression, random forest, and support vector machine
- We tuned model hyperparameters using 3-fold cross-validation within a 2/3 training set
  - We separately optimized overall accuracy (shown) and area under the ROC curve (not shown)
  - Global model parameters varied by model family
  - Individualized model parameters included patient similarity measure and cohort size
- We evaluated optimized models on the 1/3 testing set

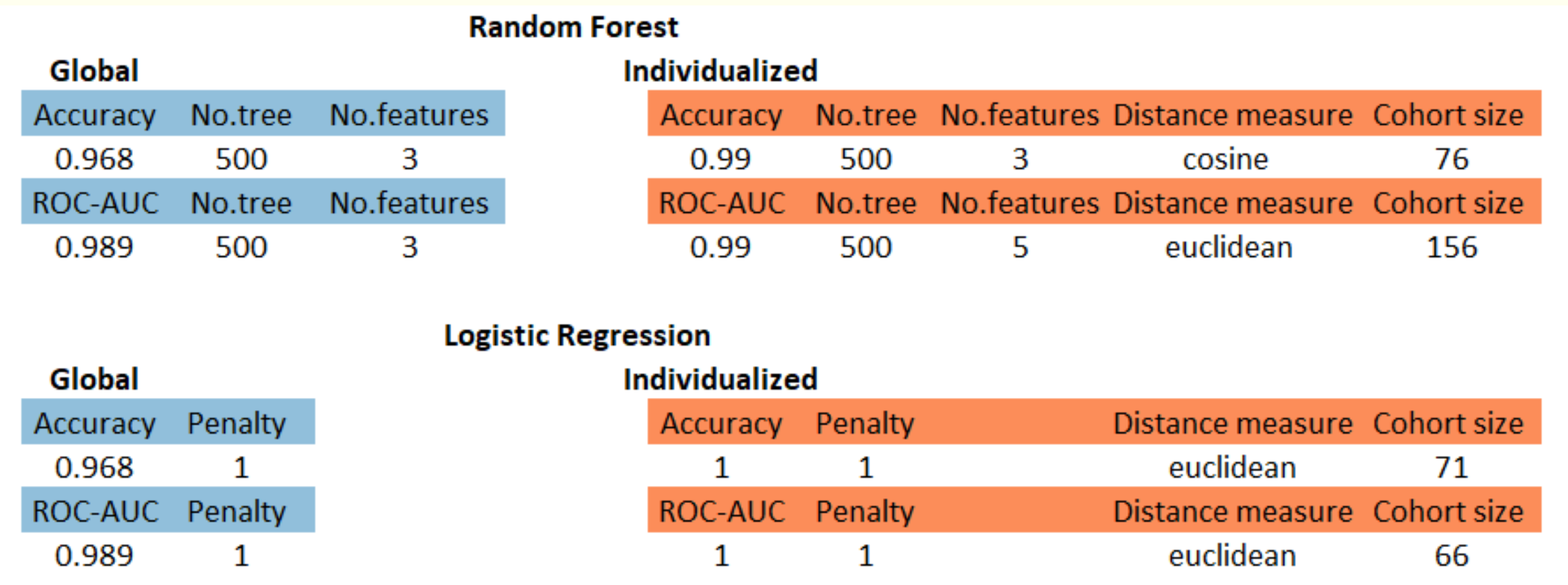
Evaluations

We compare the optimized global model performance to the optimized individualized model performance, but also allow cohort size to vary [3]:

- Size = 1: Classical case-based reasoning
- Size = training set: Global model fit

RESULTS

- The poster only shows results from model parameters optimized based on accuracy
- The left column shows results from using optimized parameters
- The right column keeps all optimized parameters fixed while varying the cohort size
- Results for optimized ROC-AUC for individualized models are mostly consistent with global models. There is also an improvement in ROC-AUC when we vary cohort sizes (figure not reported in the poster)



Random Forest

Global

Accuracy	No.tree	No.features
0.968	500	3
ROC-AUC	No.tree	No.features
0.989	500	3

Individualized

Accuracy	No.tree	No.features	Distance measure	Cohort size
0.99	500	3	cosine	76
ROC-AUC	No.tree	No.features	Distance measure	Cohort size
0.99	500	5	euclidean	156

Logistic Regression

Global

Accuracy	Penalty
0.968	1
ROC-AUC	Penalty
0.989	1

Individualized

Accuracy	Penalty	Distance measure	Cohort size
1	1	euclidean	71
ROC-AUC	Penalty	Distance measure	Cohort size
1	1	euclidean	66

Table 1. Dermatology Classification Results Comparison

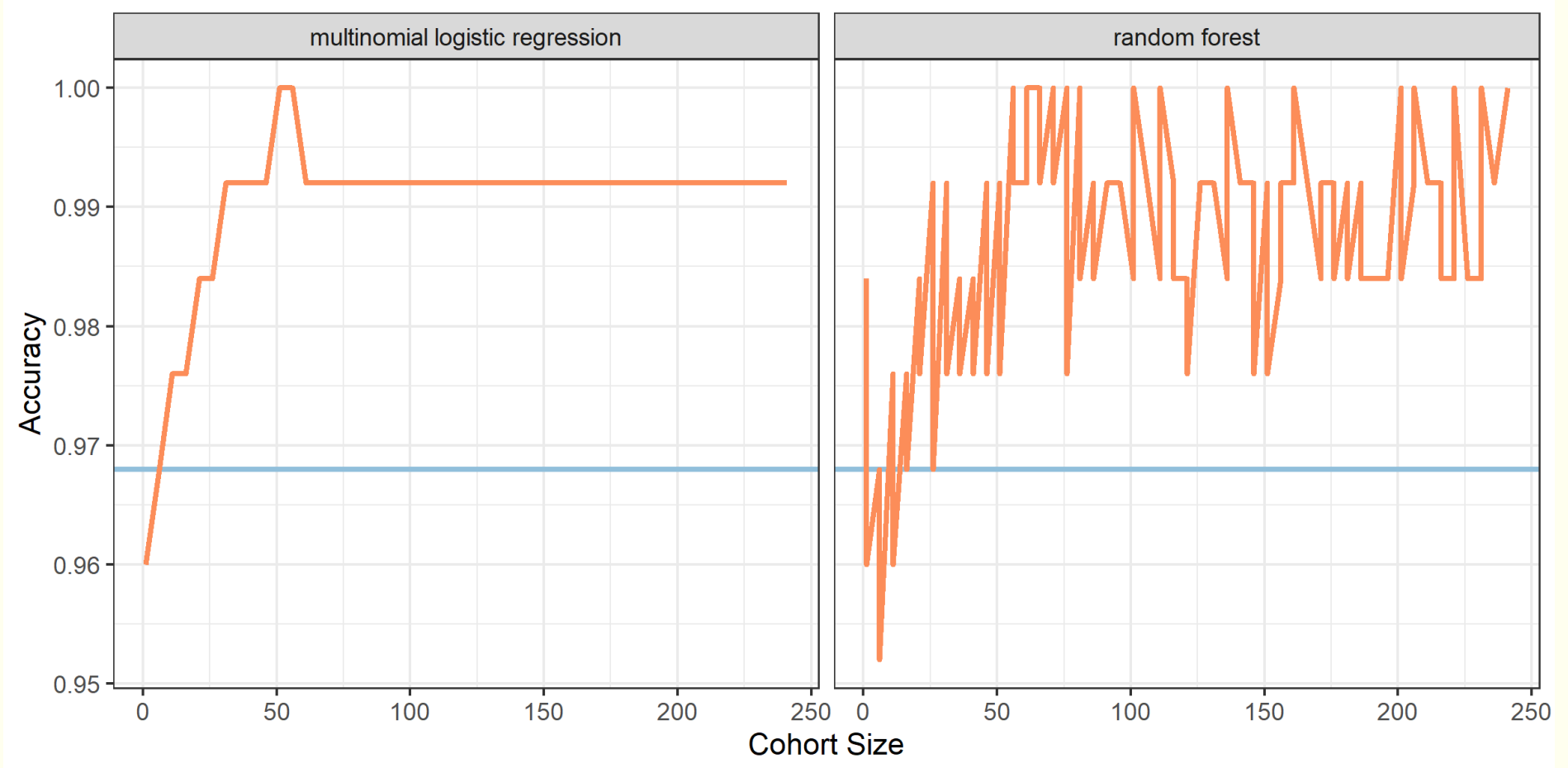
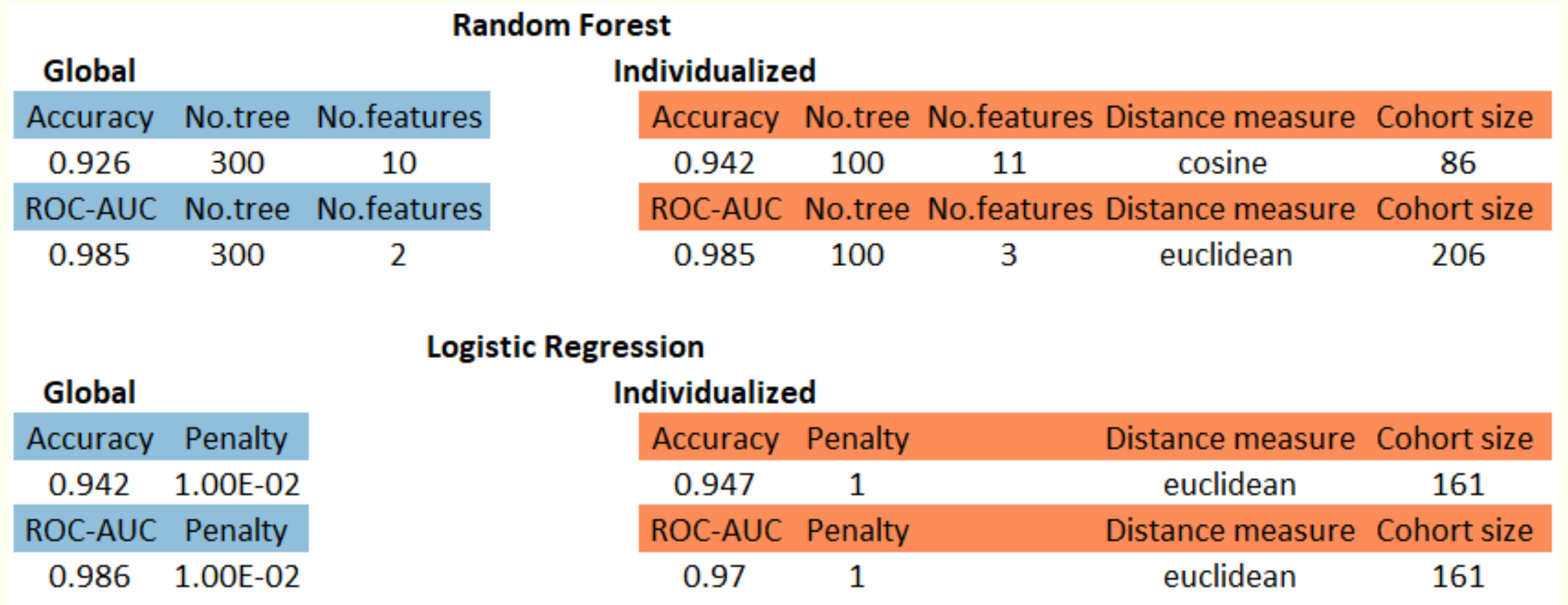


Figure 3. Dermatology Results Comparison with Different Cohort Sizes



Random Forest

Global

Accuracy	No.tree	No.features
0.926	300	10
ROC-AUC	No.tree	No.features
0.985	300	2

Individualized

Accuracy	No.tree	No.features	Distance measure	Cohort size
0.942	100	11	cosine	86
ROC-AUC	No.tree	No.features	Distance measure	Cohort size
0.985	100	3	euclidean	206

Logistic Regression

Global

Accuracy	Penalty
0.942	1.00E-02
ROC-AUC	Penalty
0.986	1.00E-02

Individualized

Accuracy	Penalty	Distance measure	Cohort size
0.947	1	euclidean	161
ROC-AUC	Penalty	Distance measure	Cohort size
0.97	1	euclidean	161

Table 2. Breast Cancer Classification Results Comparison

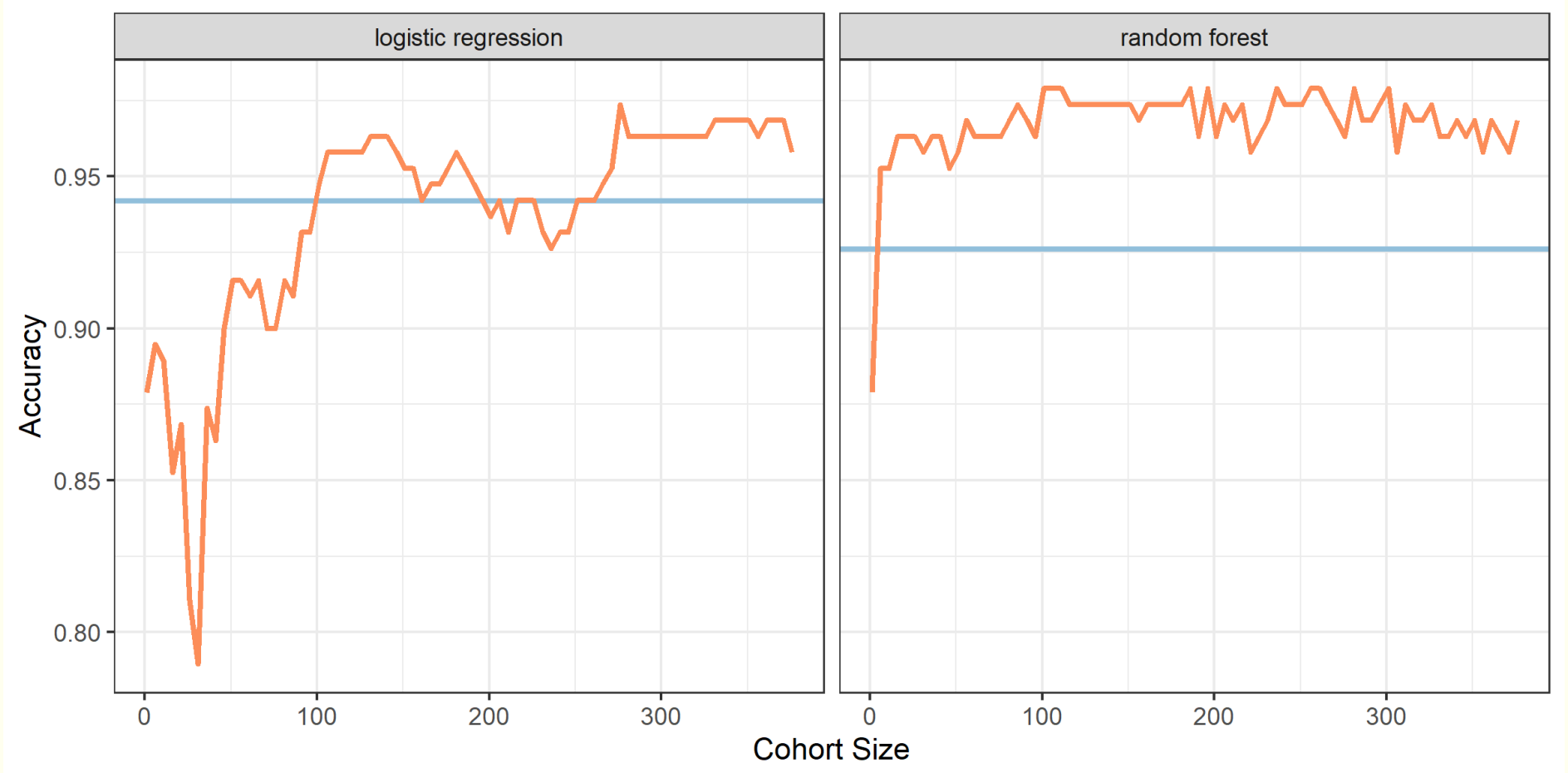
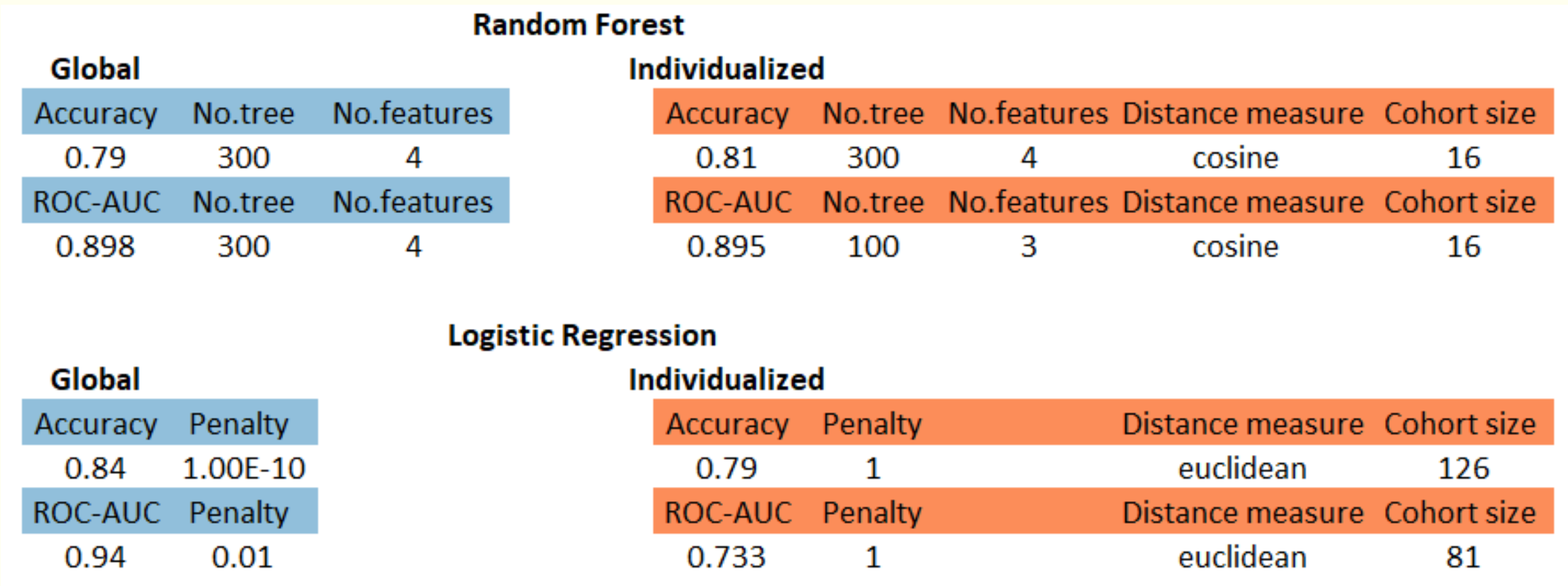


Figure 4. Breast Cancer Results Comparison with Different Cohort Sizes



Random Forest

Global

Accuracy	No.tree	No.features
0.79	300	4
ROC-AUC	No.tree	No.features
0.898	300	4

Individualized

Accuracy	No.tree	No.features	Distance measure	Cohort size
0.81	300	4	cosine	16
ROC-AUC	No.tree	No.features	Distance measure	Cohort size
0.895	100	3	cosine	16

Logistic Regression

Global

Accuracy	Penalty
0.84	1.00E-10
ROC-AUC	Penalty
0.94	0.01

Individualized

Accuracy	Penalty	Distance measure	Cohort size
0.79	1	euclidean	126
ROC-AUC	Penalty	Distance measure	Cohort size
0.733	1	euclidean	81

Table 3. Heart Disease Classification Results Comparison

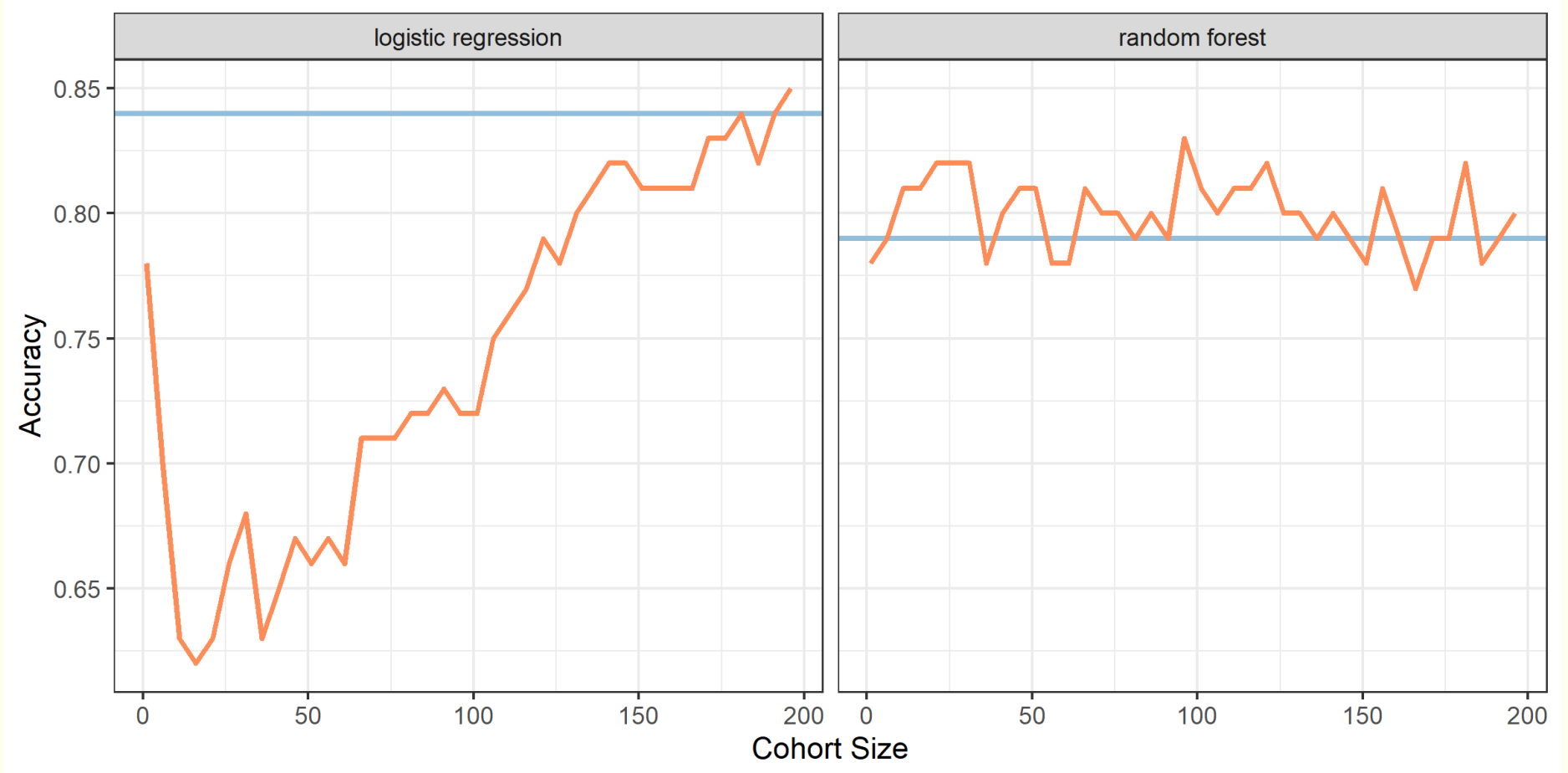


Figure 5. Heart Disease Results Comparison with Different Cohort Sizes

DISCUSSIONS

Interpretation

- We saw greater improvements in multi-class prediction than in binary prediction
- Above a certain threshold of cohort size, the individualized model outperforms the globally-fitted model
- The individualized model continues to outperform the globally-fitted model when the cohort size continues to increase

Translatability

- Individualized model has potential to be incorporated to the decision support systems
- It will be important to know which models and parameter choices yield the most informative output

Limitations

- Individualization of SVM has technical issues that have not been resolved
- Optimizing parameters for best ROC-AUC or accuracy for performance in cross-validation does not guarantee an improvement on the whole dataset

Ongoing Work

- Compare fixed-size versus similarity threshold cohorts [4]
- Compare risk factor identification & importance between models
- Identify patient subgroups that prediction can be improved the most by individualized model
- Apply individualization to predict clinical outcomes for a database of COVID-19 patients

REFERENCE

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Kuhn M, Wickham H (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.. https://www.tidymodels.org.

[3] Lee J, Maslove DM, Dubin JA (2015). Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric. PLoS ONE 10(5): e0127428. doi:10.1371/journal.pone.0127428

[4] Park, Y.-J., Kim, B.-C. and Chun, S.-H. (2006). New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis. Expert Systems, 23: 2-20. https://doi.org/10.1111/j.1468-0394.2006.00321.x