Data-driven modeling to improve pulmonary hypertension risk stratification

Team: Nikita Sivakumar, Connie Chang-Chien, Cindy Zhang, Pan Gu, Yikun Li, Yi Yang

PIs: Catherine Simpson, MD., Joseph Greenstein PhD., Casey Taylor PhD.

BIOMEDICAL ENGINEERING

INTRODUCTION & AIMS

- Pulmonary hypertension (PH) is characterized by a mean pulmonary arterial (PA) pressure greater than 20 mmHg. The inability of the right ventricle to adaptively remodel drives disease.
- Pressure-volume loops are the gold standard for assessing RV remodeling and are far superior to conventional measurements. However, this procedure is not routinely performed.
- **Aim 1.** Apply unsupervised clustering to conventional right heart catheterization (RHC) and magnetic resonance imaging (MRI) data to define distinct patient groups.
- **Aim 2.** Determine emergent metabolic phenotypes of these patient groups.
- **Aim 3.** Develop a classifier that can map conventional measurements to PV-loop groups.

AIMS 1 & 2

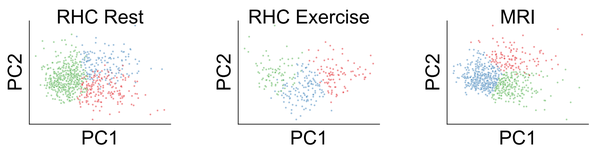

Figure 1. The PVDOMICs dataset was used to accomplish Aims 1 and 2. This dataset contains RHC, MRI, and metabolomic data for 1195 patients. For each RHC and MRI data subset, features and patients were filtered for missingness. Dimensionality reduction and unsupervised k-means clustering were performed to identify distinct groups.

AIM 3

	RHC rest	RHC exercise	MRI	All
Logistic regression	0.84	0.74	0.81	0.98
Random forest	0.88	0.76	0.73	0.98
XGBoost	0.79	0.72	0.70	0.93

Figure 2. The CALIPSO (102 subjects) dataset was used for Aim 3. PV-loop data were clustered into three groups with different survival outcomes. Classifiers were trained to predict PV-loop cluster using conventional measurements and AUCs are reported above. Using all features yielded the best results.

We applied unsupervised clustering to devise a new paradigm for pulmonary hypertension risk stratifcation.

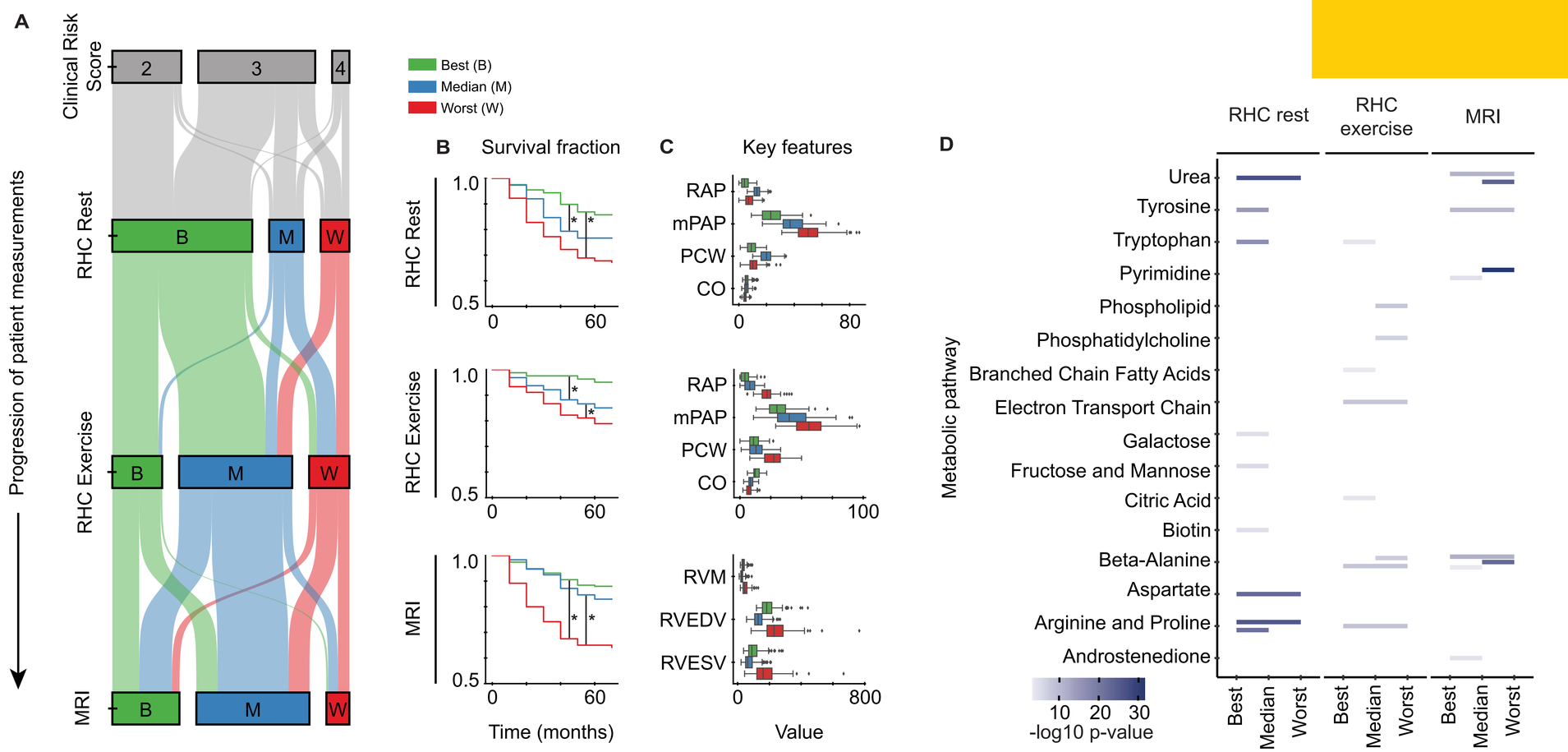

Figure 3. Unsupervised clustering of RHC rest, RHC exercise, and MRI data stratify patients into groups with distinct survival outcomes. (A) The sankey diagram describes how progressive measurements further stratify patients into distinct risk groups. **(B)** Survival v. time curves for each data subset. * = $p < 0.05$ for log-rank test. **(C)** Key features for that describe each risk group for each data subset (mPAP: Mean pulmonary arterial pressure (mmHg), PCW: Pulmonary capillary wedge pressure (mmHg), CO: Cardiac output (L/min), RVM: Right ventricle mass (g), RVEDV: Right ventricular end diastolic volume (L), RVESV: Right ventricular end systolic volume) (L). **(D)** Metabolite set enrichment analysis shows differentially expressed metabolic pathways between clusters.

We developed a classifier that can predict pressure-volume states based on conventional measurements.

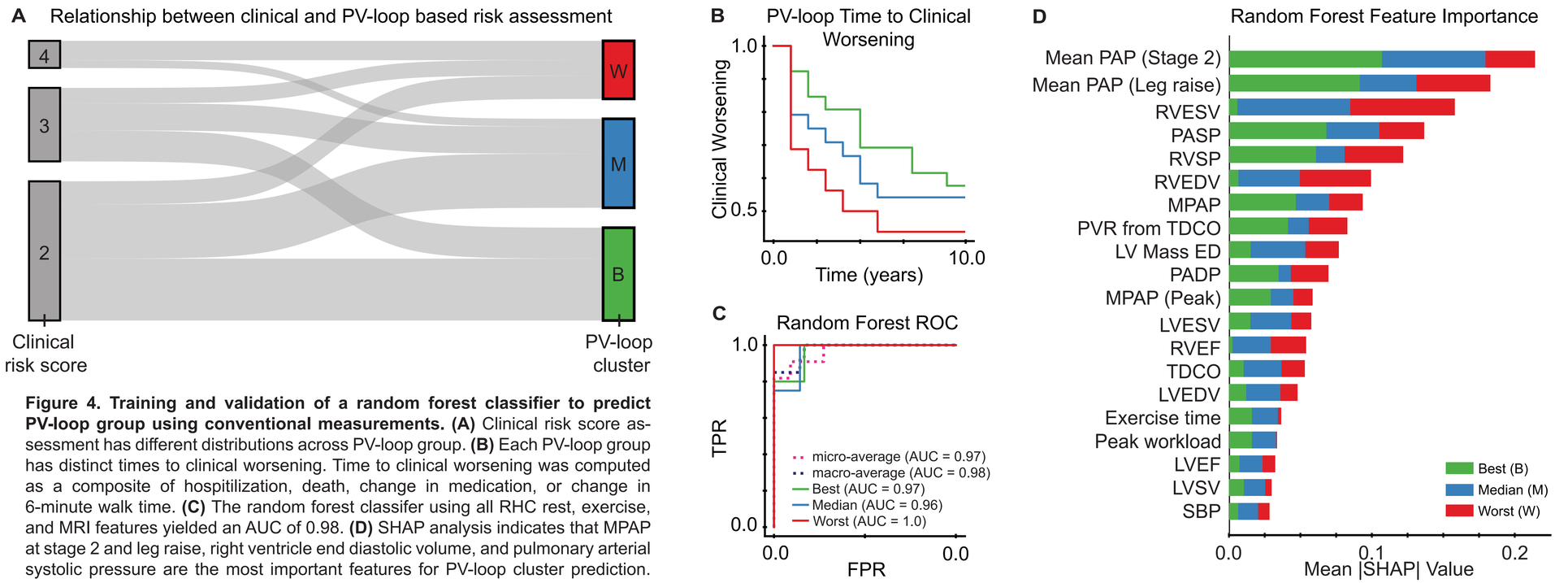

Figure 4. Training and validation of a random forest classifier to predict PV-loop group using conventional measurements. (A) Clinical risk score assessment has different distributions across PV-loop group. **(B)** Each PV-loop group has distinct times to clinical worsening. Time to clinical worsening was computed as a composite of hospitilization, death, change in medication, or change in 6-minute walk time. **(C)** The random forest classifer using all RHC rest, exercise, and MRI features yielded an AUC of 0.98. **(D)** SHAP analysis indicates that MPAP at stage 2 and leg raise, right ventricle end diastolic volume, and pulmonary arterial systolic pressure are the most important features for PV-loop cluster prediction.