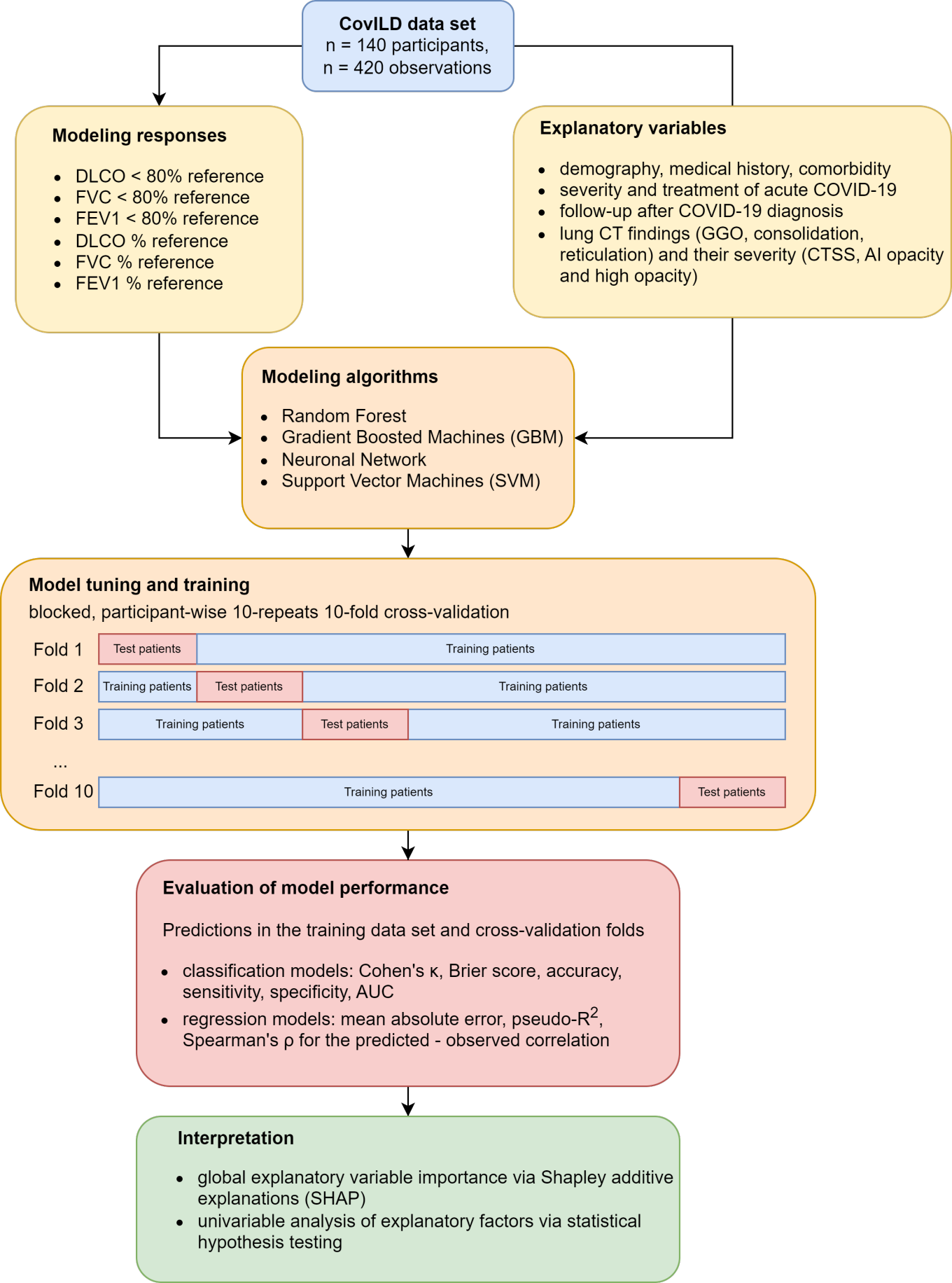
CovILD data set
n = 140 participants,
n = 420 observations
Modeling responses
DLCO < 80% reference
FVC < 80% reference
FEV1 < 80% reference
DLCO % reference
FVC % reference
FEV1 % reference
Explanatory variables
demography, medical history, comorbidity
severity and treatment of acute COVID-19
follow-up after COVID-19 diagnosis
lung CT findings (GGO, consolidation, reticulation) and their severity (CTSS, AI opacity and high opacity)
Modeling algorithms
Random Forest
Gradient Boosted Machines (GBM)
Neuronal Network
Support Vector Machines (SVM)
Model tuning and training
blocked, participant-wise 10-repeats 10-fold cross-validation
Fold 1
Test patients
Training patients
Fold 2
Training patients
Test patients
Training patients
Fold 3
Training patients
Test patients
Training patients
...
Fold 10
Training patients
Test patients
Evaluation of model performance
Predictions in the training data set and cross-validation folds
classification models: Cohen's κ, Brier score, accuracy, sensitivity, specificity, AUC
regression models: mean absolute error, pseudo-R^2, Spearman's ρ for the predicted - observed correlation
Interpretation
global explanatory variable importance via Shapley additive explanations (SHAP)
univariable analysis of explanatory factors via statistical hypothesis testing