Research | Open Access | Published: 25 May 2022

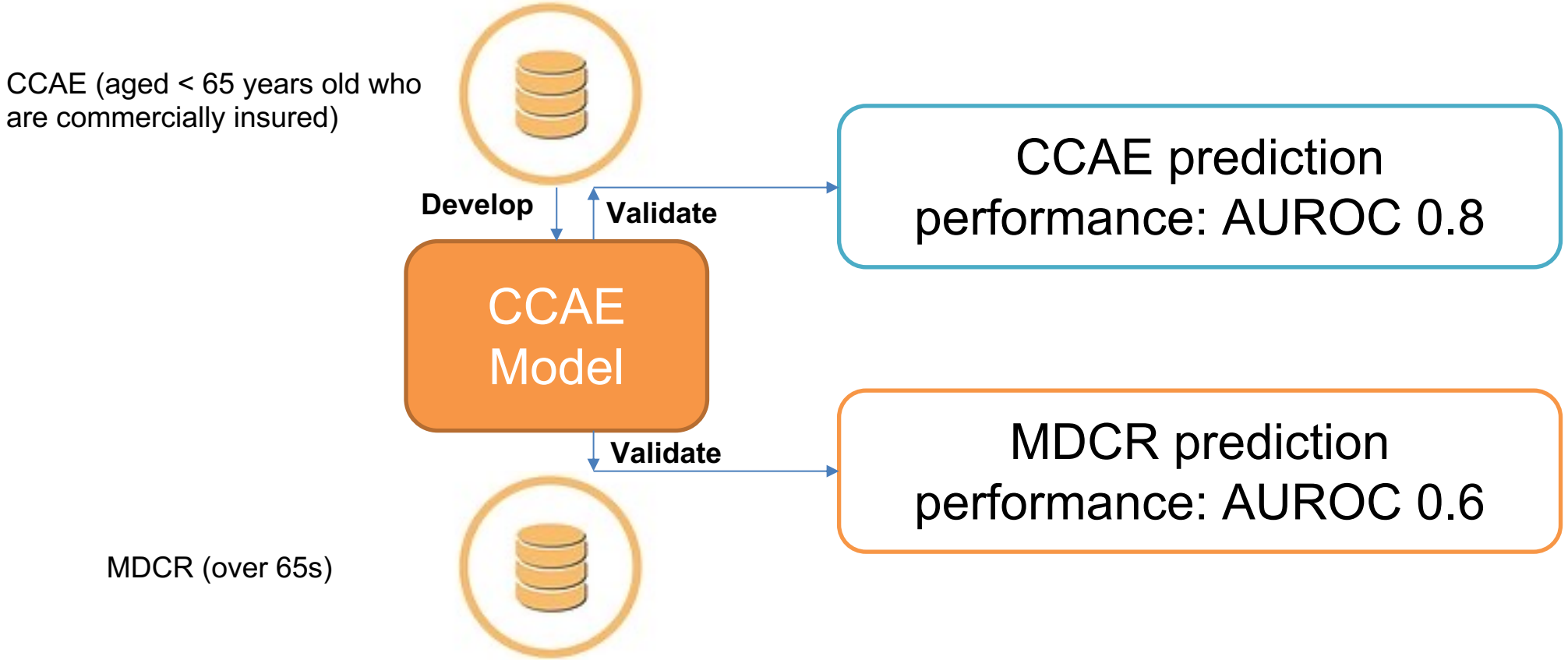
Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability

Jenna Marie Reps ✉, Ross D. Williams, Martijn J. Schuemie, Patrick B. Ryan & Peter R. Rijnbeek

BMC Medical Informatics and Decision Making **22**, Article number: 142 (2022) | Cite this article

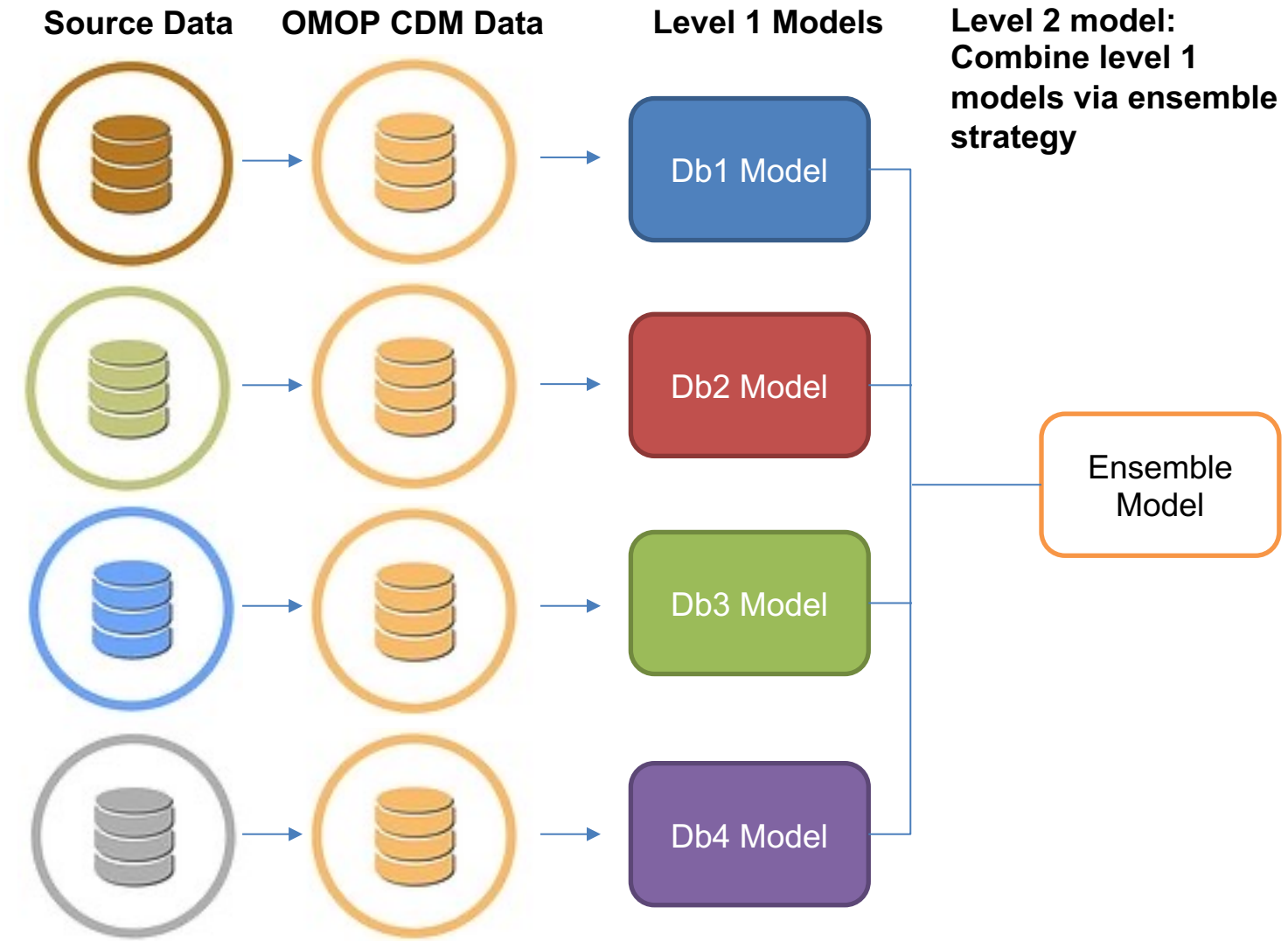
https://link.springer.com/article/10.1186/s12911-022-01879-6

Motivation: Models sometimes do not transport across databases



Question: If a model does not transport across databases – will it transport into a clinical setting?

In this paper we trained models across different databases and then used ensemble learning to combine them…



1. Level 1 models developed using PatientLevelPrediction package.

Same:

- Target population
- Outcome
- Candidate covariates Machine learning model

Different:

- Development database.

2. Level 2 model combines the level 1 models using different strategies:

- Fusion
- Stacker
- Mixture of experts

Ensemble strategies:

Fusions:

- Weighted combination of level 1 models' predicted risk
- Simple to implement – can just take the mean predicted risk across level 1 models (uniform ensemble)

Mixture of experts:

- Picks one level 1 model's predicted risk per patient based on some rule

Stacker:

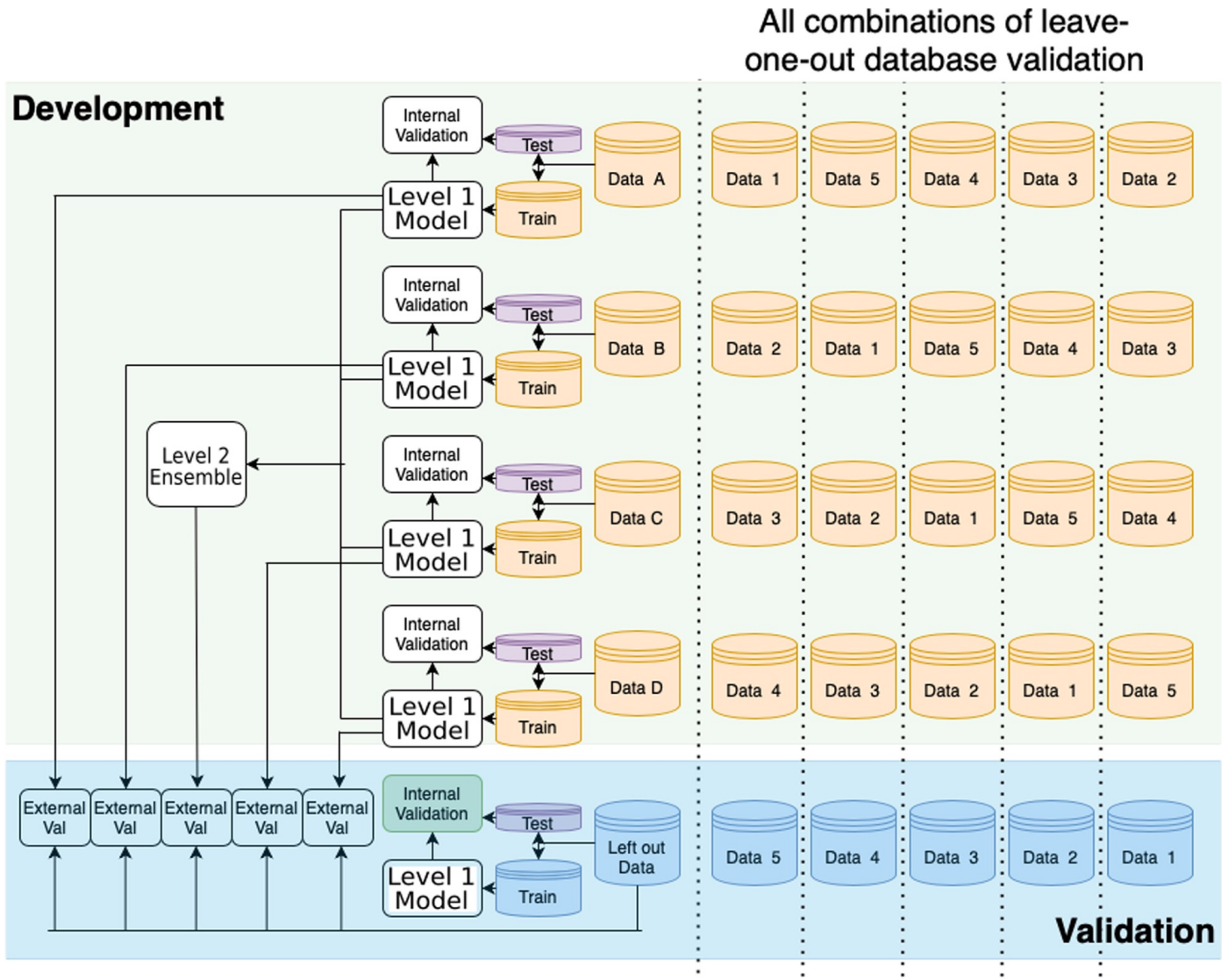
- Train a supervised model that uses the level 1 models' predictions as features
- Requires more labelled data to learn from (in this paper we used some data from the held-out database)

Methods: We used a leave-one-database out validation

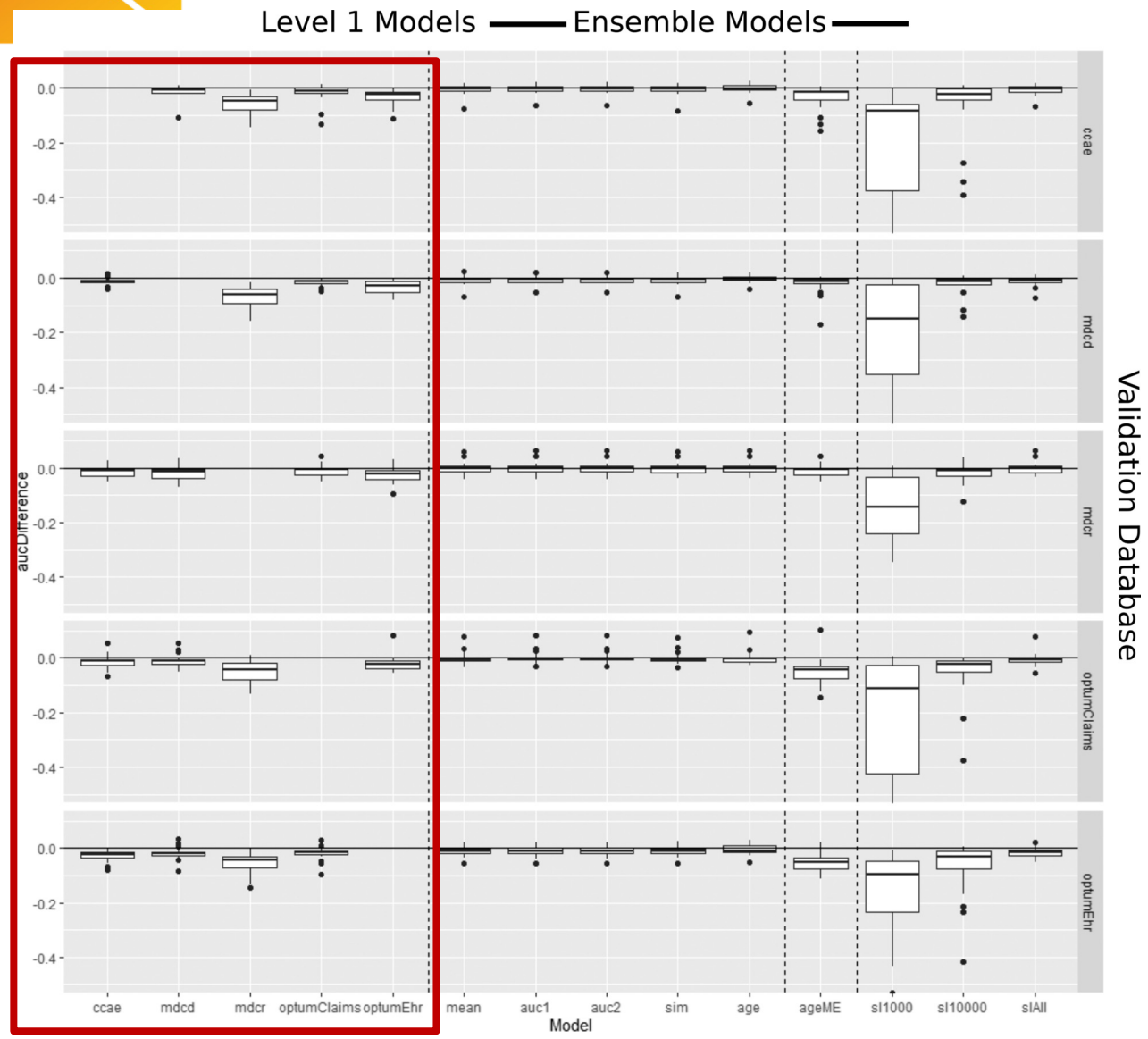
Databases:
CCAE, MDCR, MDCD, Optum claims, Optum EHR

Models:
Level 1 database models, plus various ensembles (combining these level 1 database)

Performance:
discrimination (AUROC) and calibration (mean predicted risk vs observed risk)



Results: Discrimination



Transportability

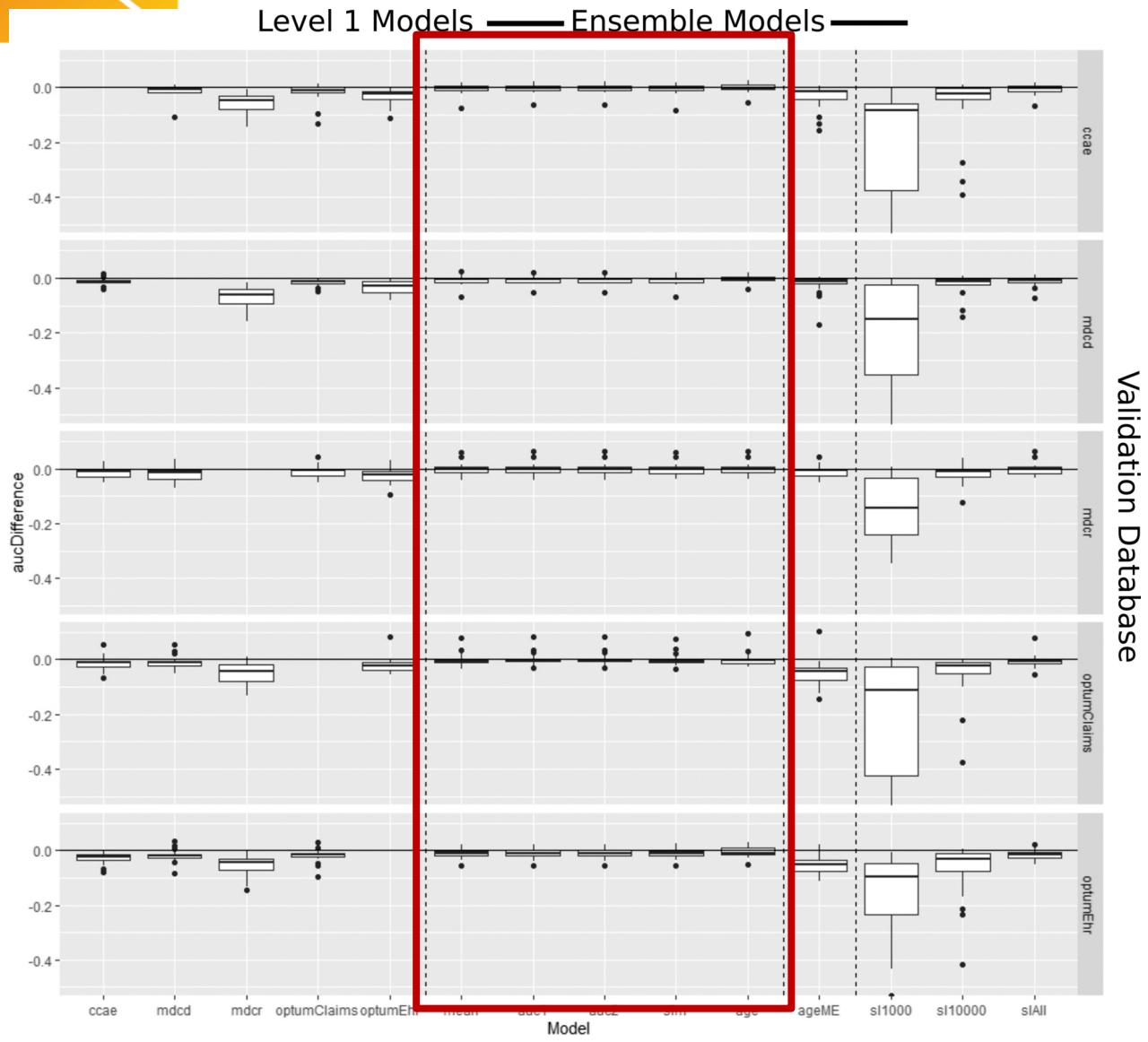
Above 0 —— 0 —— Below 0

Good **Worse**

Level 1 models (model trained on one database):

- Mostly slightly below 0 so s slightly worse transportabiltiy

Results: Discrimination



Transportability

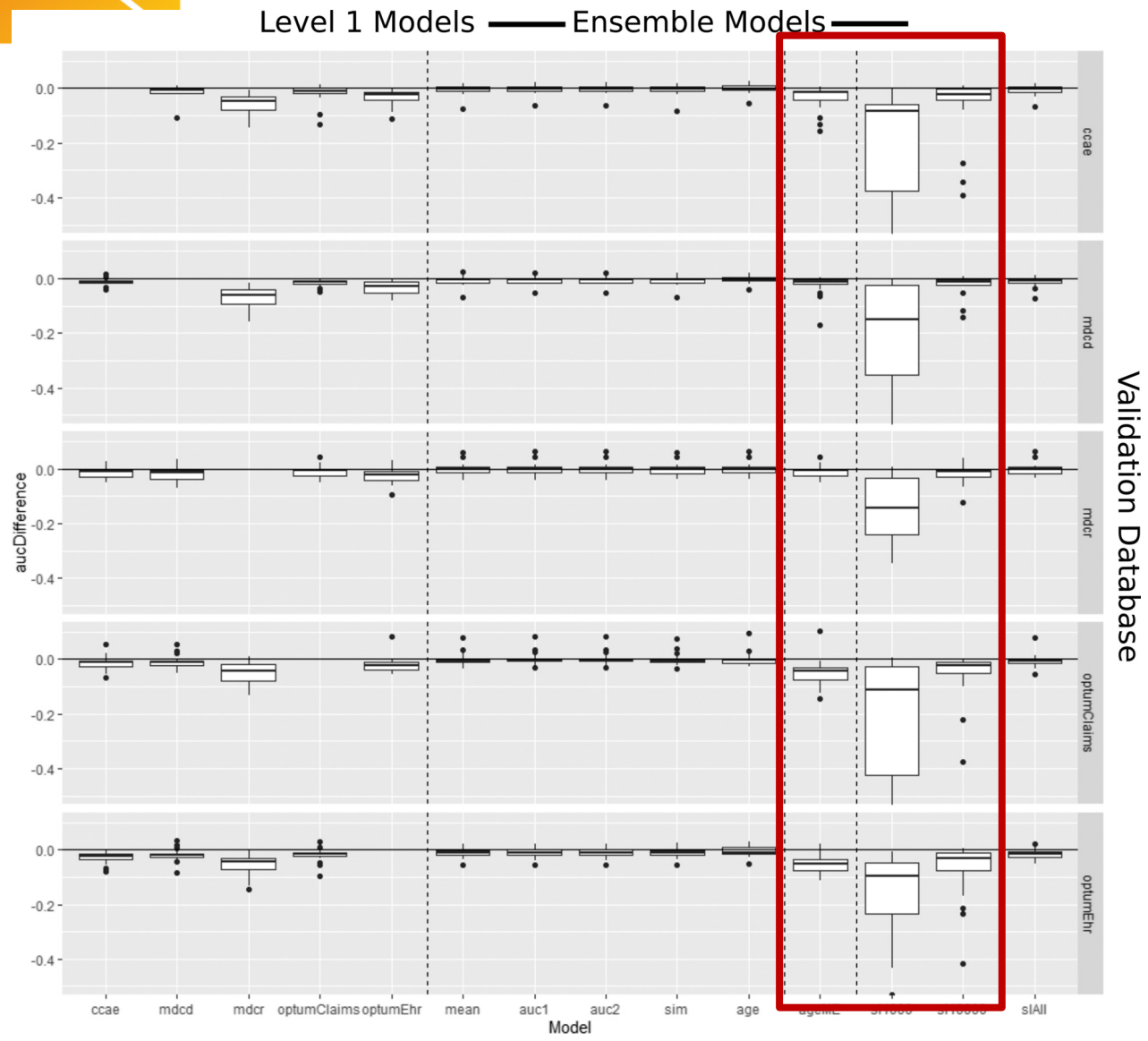
Above 0 —— 0 —— Below 0

Excellent **Good** **Worse**

Fusion ensembles (weighted combination of level 1 models predictions):

- Mostly around 0 so **good transportability** in terms of discrimination
- Better than the level 1 models.

Results: Discrimination



Transportability

Above 0 —— 0 —— Below 0

Excellent **Good** **Poor**

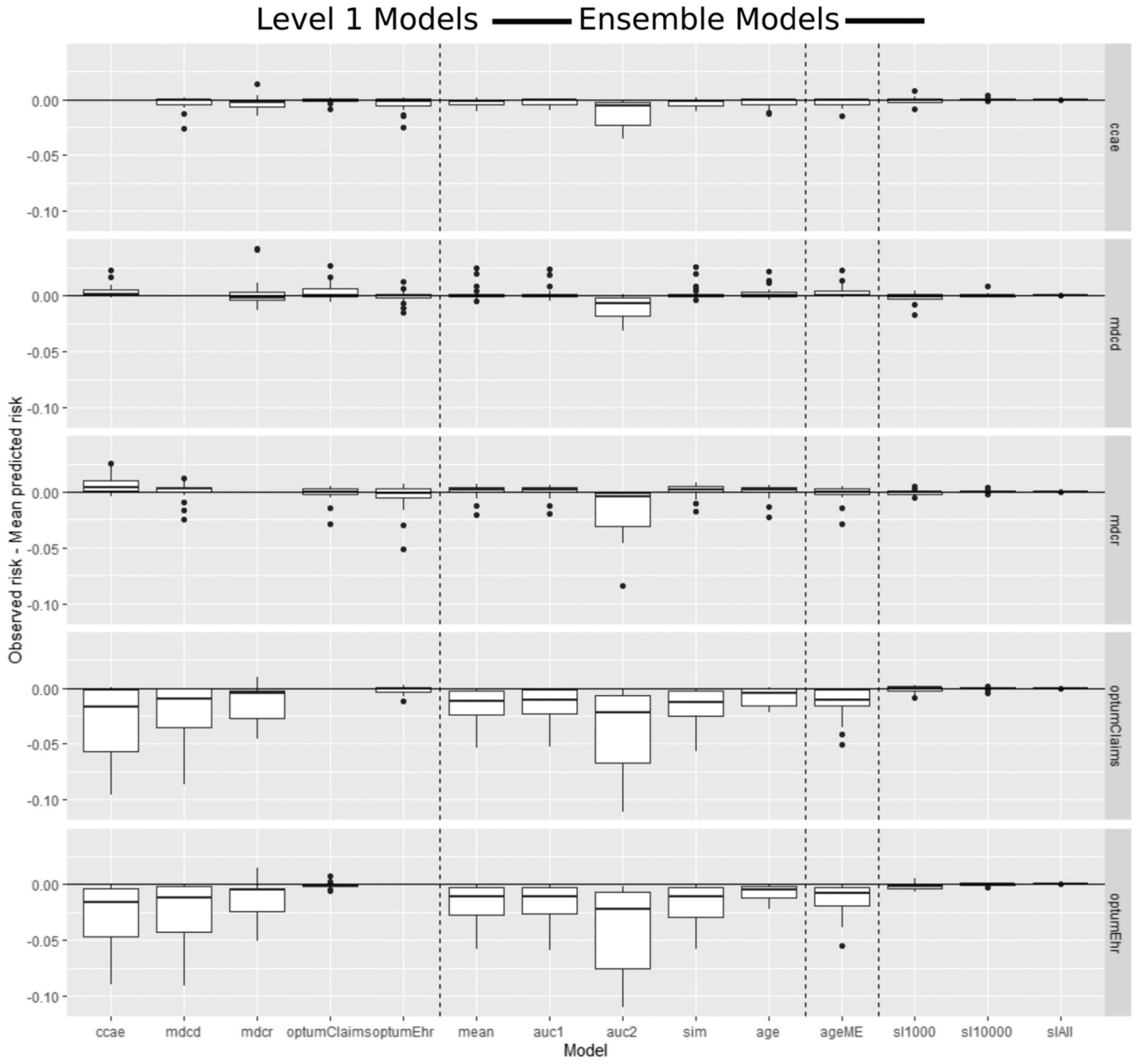
Other ensembles (mixture of experts/stacker):

- Some had very poor transportability

Results: Calibration

Y-axis = difference between mean predicted risk and observed risk (so anything above or below 0 means poor calibration)

All models are poorly calibrated (except the stackers that were effectively recalibrated)



We now have an R package for ensemble learning

EnsemblePatientLevelPrediction

R-CMD-check passing codecov 91%

EnsemblePatientLevelPrediction is part of HADES.

Introduction

EnsemblePatientLevelPrediction is an R package for building and validating ensemble patient-level predictive models using data in the OMOP Common Data Model format. The package expands the OHDSI R PatientLevelPrediction package to enable ensemble learning.

In our study here we found that combining models developed using different databases resulted in models that had better discrimination performance compared to the level 1 models (single database) when transported to new data. However, calibration was poor. This has prompted the EnsemblePatientLevelPrediction package where users can combine models developed on the same database or models developed on different databases.

User Documentation

Vignette: EnsemblePatientLevelPrediction

Website: Documentation can be found on the package website.

Package manual: EnsemblePatientLevelPrediction.pdf

https://github.com/OHDSI/EnsemblePatientLevelPrediction

Questions?

Email: jreps@its.jnj.com

The different ensemble strategies: **Uniform Fusion**

PATIENT	Db1 Model prediction	Db2 Model prediction	Db3 Model prediction	Db4 Model prediction
1	40%	50%	30%	80%
2	8%	1%	1%	2%
…				

Ensemble
= (40+50+30+80)/4 = 50%
= (8+1+1+2)/4 = 3%

The different ensemble strategies: **Weighted Fusion**

X1 = 0.2 X2 = 0.3 X3 = 0.4 X4 = 0.1

PATIENT	Db1 Model prediction	Db2 Model prediction	Db3 Model prediction	Db4 Model prediction	Ensemble
1	40%	50%	30%	80%	= (40*x1+50*x2+30*x3+80*x4)/4
2	8%	1%	1%	2%	= (8*x1+1*x2+1*x3+2*x4)/4
…					

Weights based on:

- AUROC in test set
- Data similarity
- Patient similarity

The different ensemble strategies: **Mixture of expert**

PATIENT	Db1 Model prediction	Db2 Model prediction	Db3 Model prediction	Db4 Model prediction
1	40%	**50%**	30%	80%
2	**8%**	1%	1%	2%
…				

Ensemble
= 50%
= 8%

Base model selected on:

- Patient age similarity

The different ensemble strategies: **Stacker**

PATIENT	Db1 Model prediction	Db2 Model prediction	Db3 Model prediction	Db4 Model prediction
1	40%	50%	30%	80%
2	8%	1%	1%	2%
…				

Strategy: Learn another model that uses the Db1-Db4 predictions as predictors (this requires some more labelled data)