

Curated phenotypes made federated training much faster without materially changing discrimination.

Feasibility of Fully Data-Driven Federated Learning on Large Observational Health Data

Background: We simulated 5 sites within IPCI and compared federated and pooled L1 penalized logistic models predicting risk of dementia, lung cancer, or readmission, using either age, sex, and last-year condition indicators, or age, sex, and 63 curated phenotypes as features.

Federated and global models achieved very similar AUCs across dementia, lung cancer, and readmission.

Nested-CV AUCs

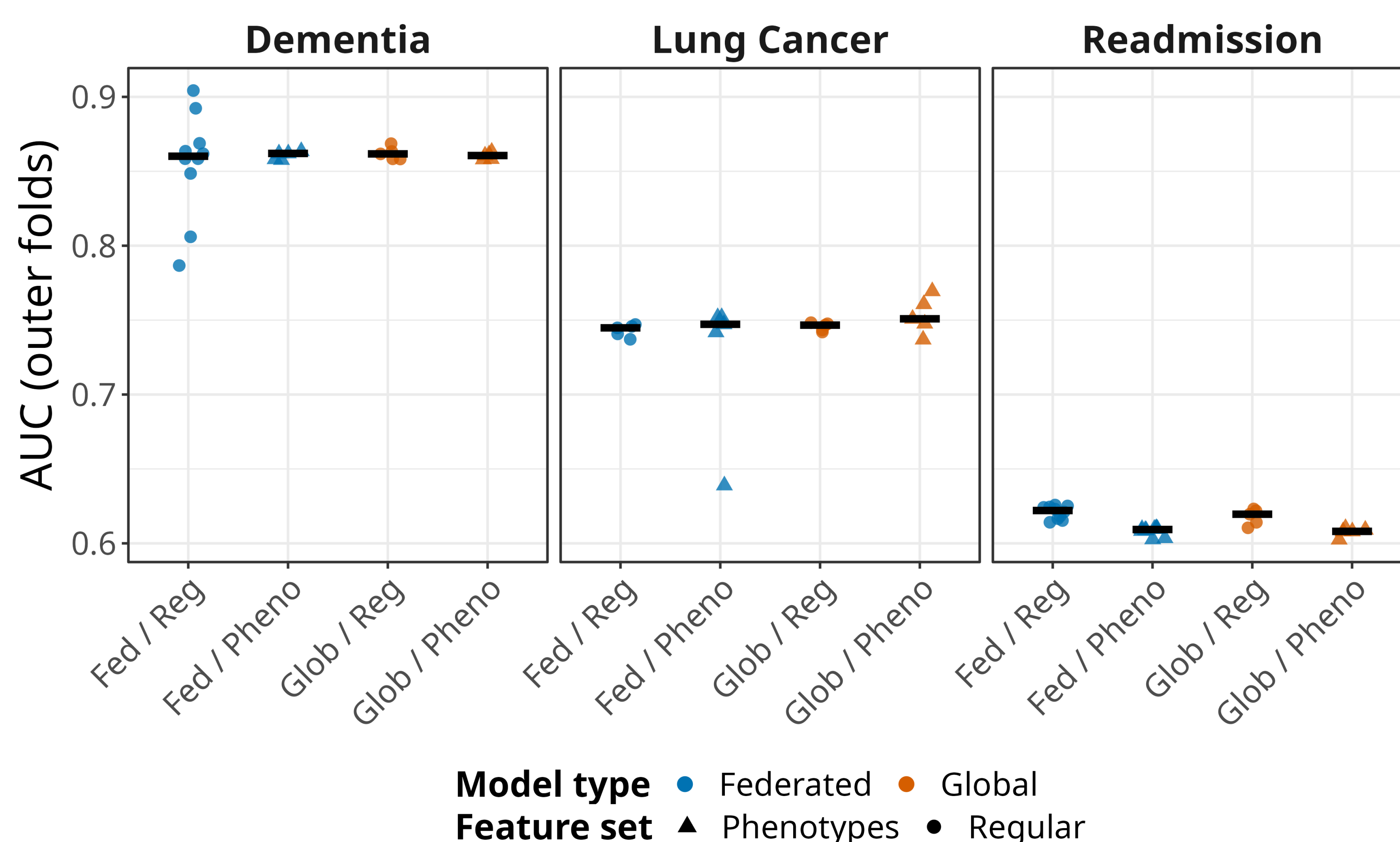


Figure 1: Outer-fold AUCs by task and model.

Study design

Study setup

- IPCI Netherlands
 - split into 5 simulated sites
- 3 tasks
 - dementia, lung cancer, readmission
- federated vs pooled/global models
 - Federated training used dual averaging

What is compared

- Regular features
 - age, sex, and last-year conditions
- Phenotypes
 - age, sex and 63 curated predictors
- 5-fold nested CV

Using 63 curated phenotypes reduced runtime sharply, especially for federated training.

Training runtime

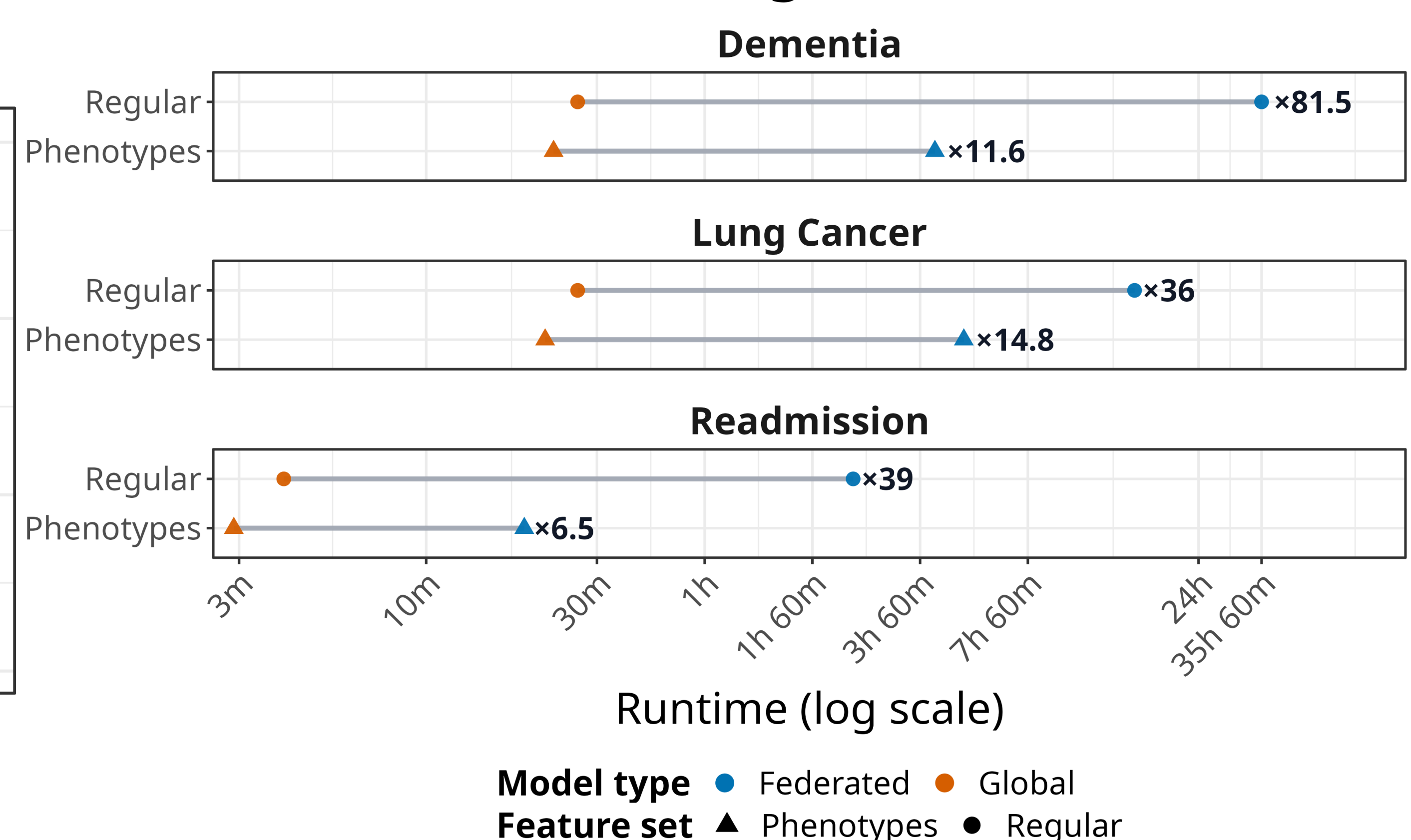
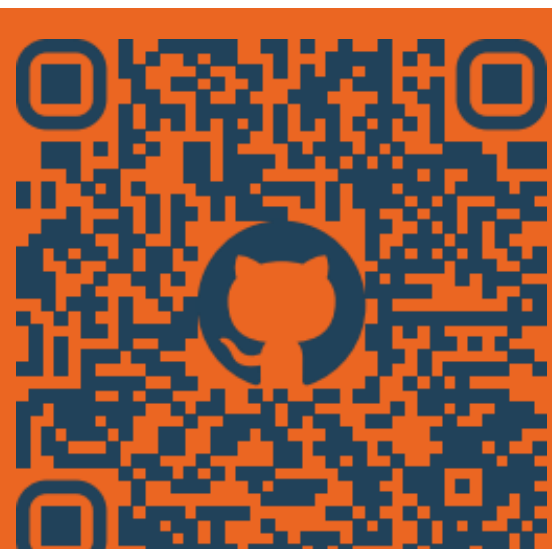


Figure 2: Log-scale runtime; labels show slowdown.

Limitation: Runtimes come from a simulated single-machine setup and exclude network or secure-computation overhead.



Egill A. Fridgeirsson¹, Jenna M. Reps^{1,2}

e.fridgeirsson@erasmusmc.nl

1: Department of Medical Informatics Erasmus University Medical Center, Rotterdam, the Netherlands
2: JnJ, Raritan, New Jersey, USA

