Extensive Benchmarking of a Method that Estimates External Model Performance from Limited Statistical Characteristics

Supplementary Information

Tal El-Hay[1], Jenna M Reps[2], Chen Yanover[1]

[1]KI Research Institute, Kfar Malal, Israel.
[2]Janssen Research and Development, Raritan, NJ, USA.

Supplementary Note 1 Feature sets used for weighting

The feature sets used for weighting do not have to be the same as the model's features. The dependence of external evaluation performance on the feature sets used for weighting was tested for logistic regression models with different feature sets.

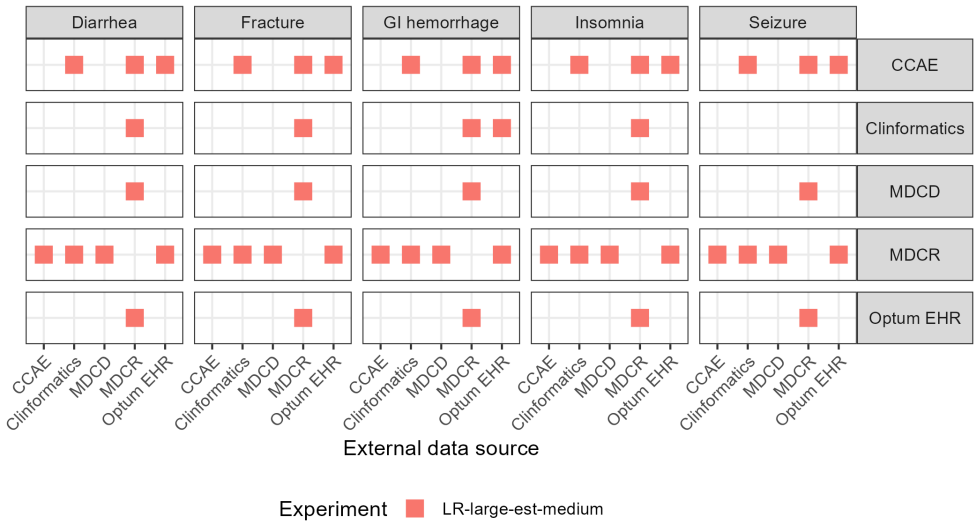
Supplementary Note 2 Large feature set model

Performance estimation of the linear model trained with the large feature set was tested with three different features sets in the weighting step:

1. Important features as defined in the main benchmark
2. The medium feature set of 64 phenotype features
3. Top 100: the 100 features with the largest absolute values of model coefficients

Supplementary Figure 1 shows the cases in which the evaluation pipeline did not provide results.

Results were missing only when using the phenotype predictors (medium feature set). Almost all cases that involved Medicare as an internal or as an external data source were not evaluated (pre-diagnostics failed). Additionally, when the internal dataset was CCAE, the estimation algorithm did not provide results for Optum's external datasets. This was most likely due the inclusion of age (in years) in the summary statistics for the phenotype predictors. The Optum® de-identified Electronic Health Record data source and Optum's de-identified Clinformatics® Data Mart Database contain patients aged 65+ but CCAE does not, so it was impossible to find a weighted subset of CCAE that matched the summary statistics of the external resources when age in years was included. Performance is shown in Supplementary Figure 2.

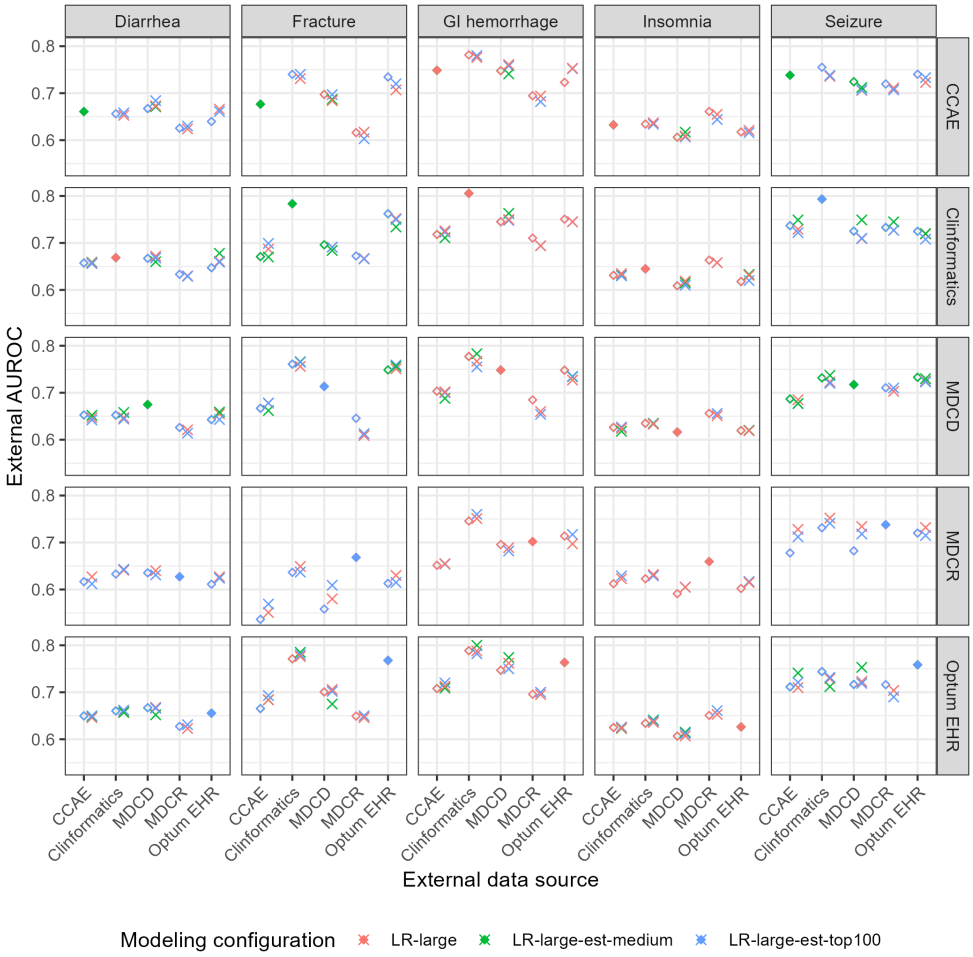


Supplementary Figure 1 Cases that lack an estimation of large feature set model performance. The *significant features* and *Top 100* configurations had estimations for all cases. The *medium feature set* configuration was the only one in which some cases lacked an estimation. Such cases are represented by red squares. Columns correspond to outcomes and rows to internal data source.

Supplementary Note 3 Small feature set model

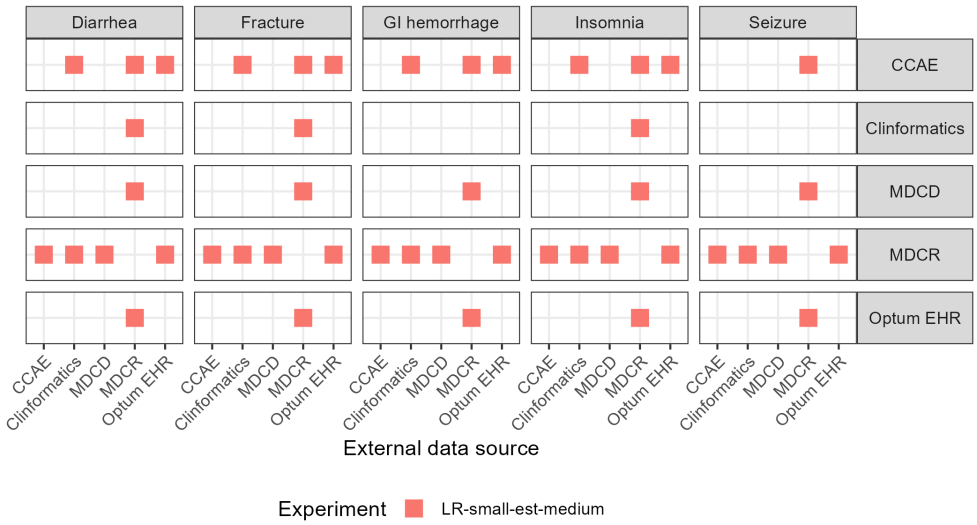
Performance estimation of the linear model trained with the small feature set (that is, only age and sex features) was tested with three different feature sets in the weighting step:

1. The small feature set
2. Age, sex and interactions between them
3. The medium feature set of 64 phenotype features



Supplementary Figure 2 Performance estimation for the large linear model using three different feature sets. Filled and empty diamonds correspond to internal and external test performance, respectively; estimated AUROC values are marked by ×. Columns correspond to outcomes and rows to internal data source. LR-large, estimation with important features of the large linear model. LR-large-est-medium, estimation with the medium feature set. LR-large-est-top100, estimation with the top 100 features.

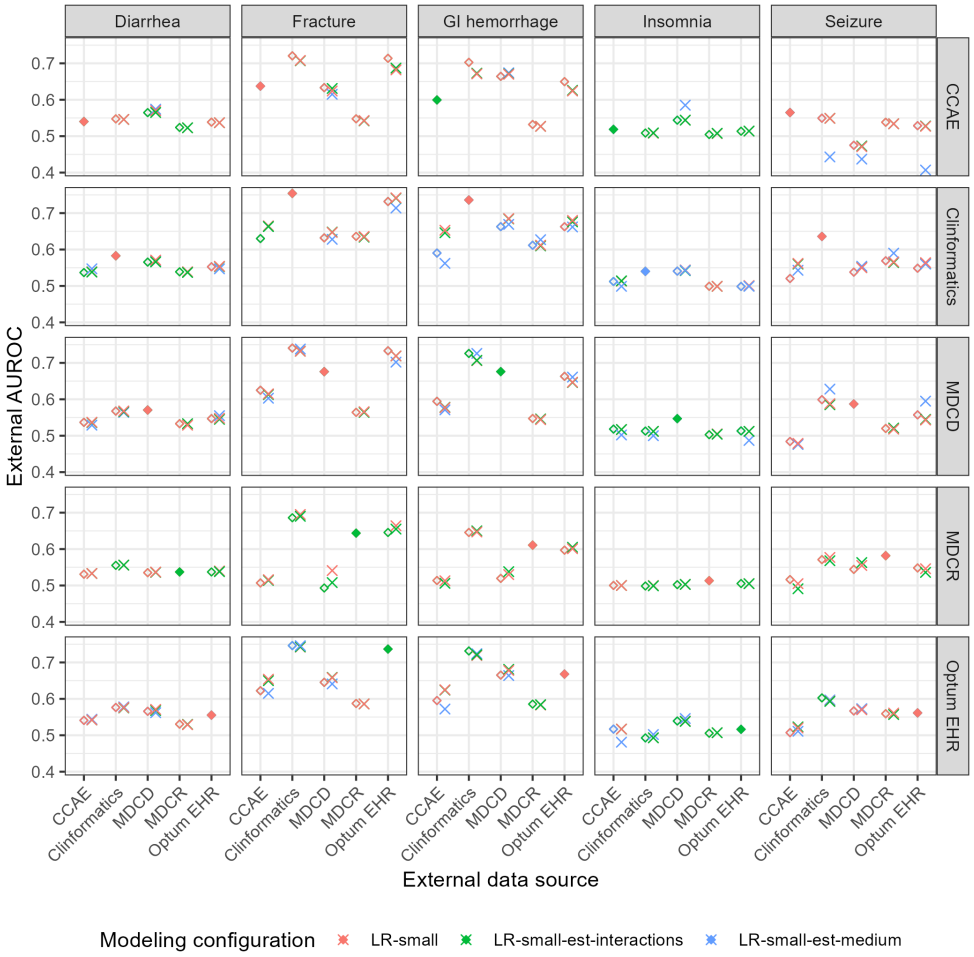
The results for the small model showed a similar dependency on the selected feature set as shown for the large model. In other words, estimations may be lacking when the feature sets are not related to the model; see Supplementary Figures 3 and 4.



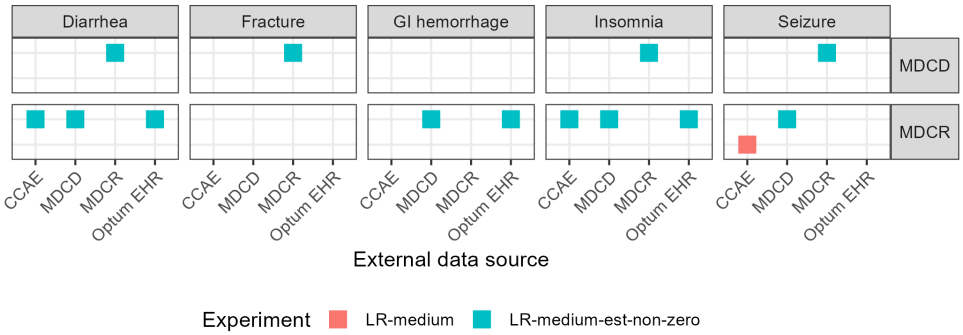
Supplementary Figure 3 Cases that lack an estimation of small models' performance. Estimations that used sex and age features, with or without interactions, were available in all cases. Estimations that used the medium feature sets lacked in some cases, shown in red squares. Columns correspond to outcomes and rows to internal data source.

Supplementary Note 4 Medium feature set model

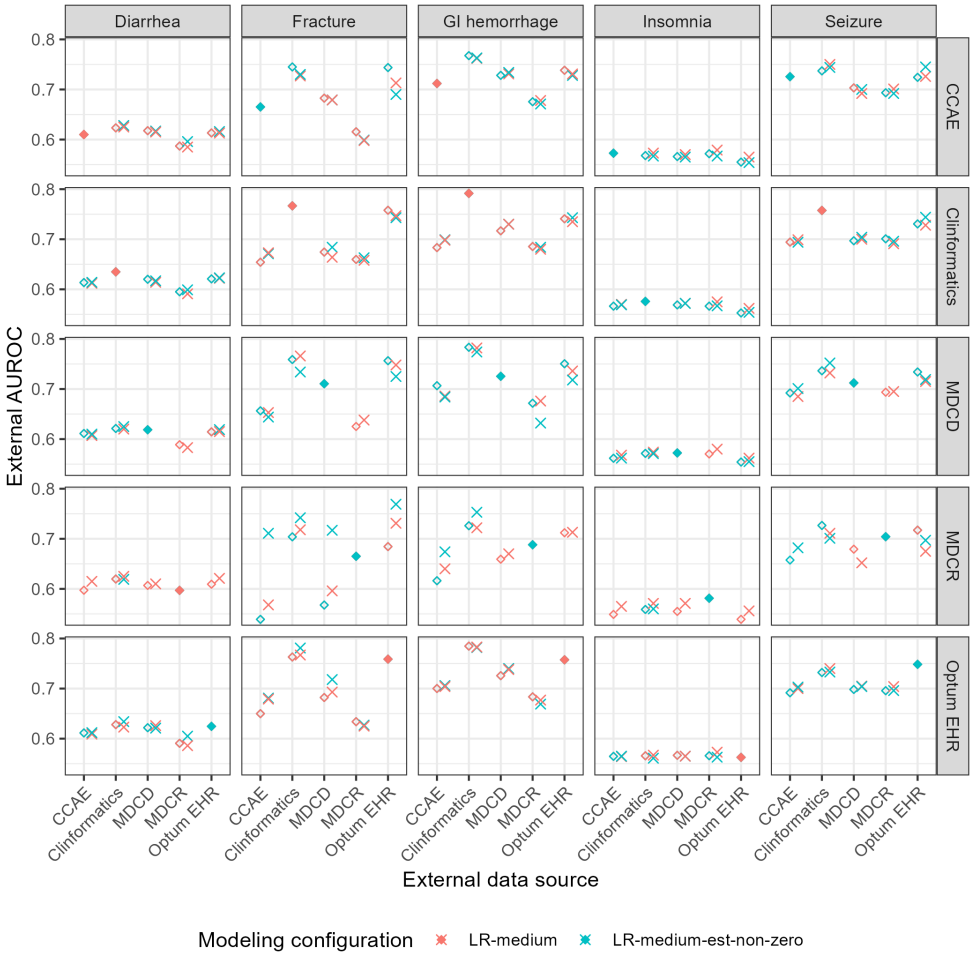
In the medium linear model we compared the algorithm when using features with sufficiently high coefficients (absolute value ≥ 0.1) versus all features with non-zero coefficients, see Supplementary Figures 5 and 6.



Supplementary Figure 4 Performance estimation for the small model using three different feature sets. Filled and empty diamonds correspond to internal and external test performance, respectively; estimated AUROC values are marked by ×. Columns correspond to outcomes and rows to internal data source. LR-small, estimation using the small feature set. LR-small-est-interactions, estimation using age sex features with interaction terms. LR-small-est-medium, estimation with the medium feature set.



Supplementary Figure 5 Cases that lack an estimation of the medium feature set linear models' performance. Columns correspond to outcomes and rows to internal data source. LR-medium, re-weighting using logistic regression coefficient with absolute value ≥ 0.1. LR-medium-est-non-zero, using all the features that were not regularized to zero by the penalty term of the training procedure.



Supplementary Figure 6 Performance estimation in the medium feature set linear model using two different feature sets. Filled and empty diamonds correspond to internal and external test performance, respectively; estimated AUROC values are marked by $\times$. Columns correspond to outcomes and rows to internal data source. LR-medium, re-weighting using logistic regression coefficient with absolute value ≥ 0.1. LR-medium-est-non-zero, using all the features that were not regularized to zero by the penalty term of the training procedure.