External validation using clinical domain knowledge from the SNOMED medical terms hierarchy

LH John[1], EA Fridgeirsson[1], JM Reps[2], JA Kors[1], PR Rijnbeek[1]
[1]Erasmus University Medical Center, [2]Janssen Research and Development

Background

External validation is crucial for ensuring the reliability of prediction models on new data. However, performance often declines during external validation due to database heterogeneity caused by variations in record collection, regulatory guidelines, and database purposes. [1]

<u>Use Case:</u> Figure 1 depicts a hypothetical model developed on the Integrated Primary Care and Information, a Dutch GP database, with predictors *Heart failure*, *Depression*, and *COPD*, which cannot be applied to a patient from an external database who has slightly different diagnoses. However, considering the contextual similarity, a medical expert may have been able to apply the model based on clinical domain knowledge.

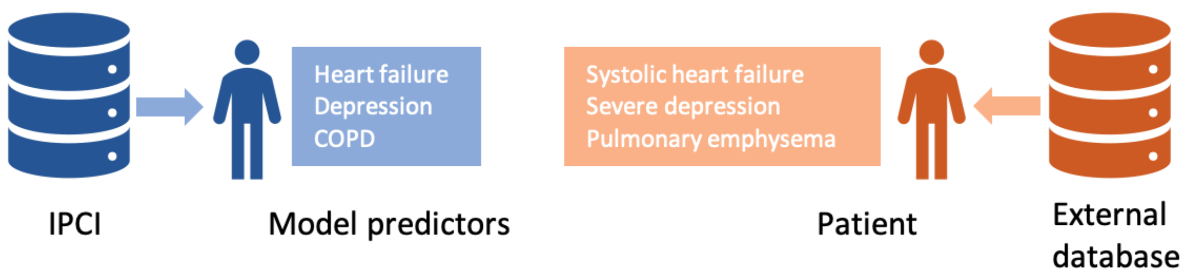

Figure 1. Incompatible model and patient record due to database heterogeneity.

This work aims to utilizes embeddings to approximate clinical concepts, specifically in the context of predicting dementia in persons aged 55-85 in the next five years. This approach may enable external validation of a model even when an exact match for predictors is not found in a patient's record.

Methods

Clinical domain knowledge is encoded in our vocabulary hierarchies. For example, SNOMED provides over one million ancestor-descendant relationships. Figure 2 shows a subset of 177 SNOMED relationships with the ancestor concept *Clinical finding* as tree root. In this work we embed the SNOMED hierarchy, to obtain a latent space in which items that resemble one another are positioned closer to each other, which will allow us to approximate missing concepts.

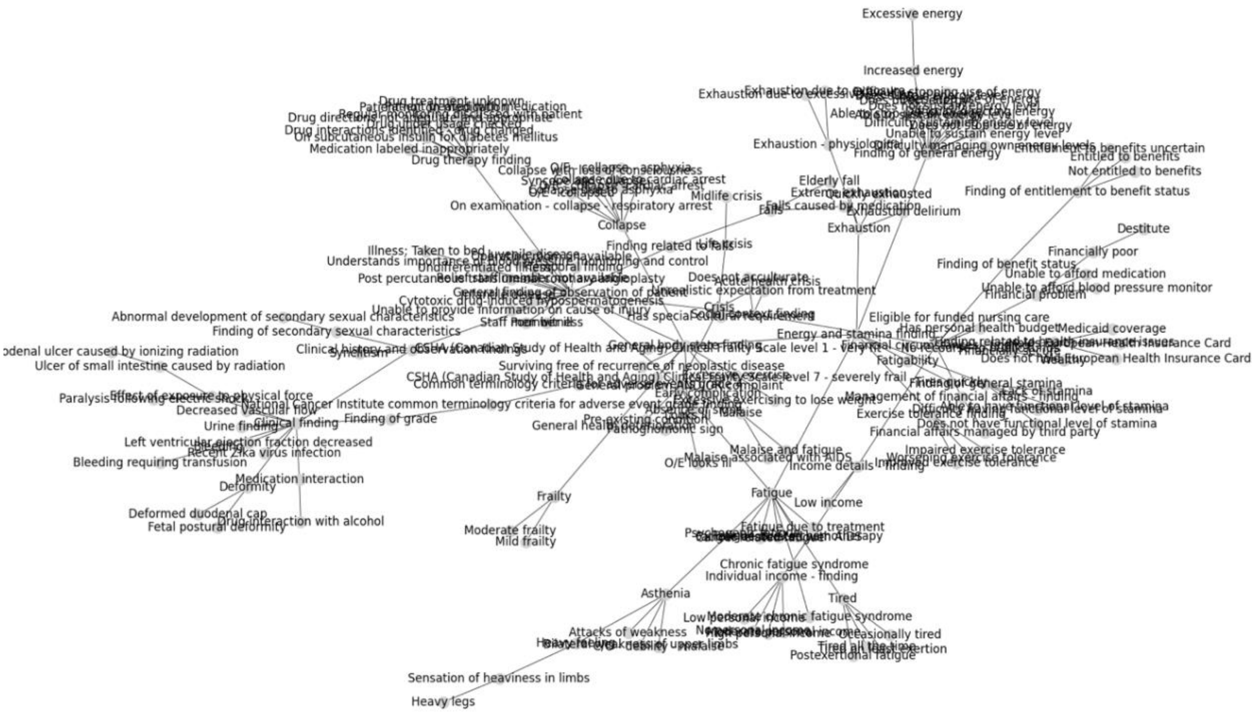

Figure 2. Subset of SNOMED medical terms hierarchy with the concept Clinical finding as the root.

Nickel & Kiela introduced an efficient method to embed hierarchical data, such as the SNOMED hierarchy, into a lower-dimensional manifold [2]. Hierarchical data follows a tree structure. The number of descendants exponentially increases with distance from the root. To address the limitation of growing hierarchical data, which can exceed the available Euclidean space in Euclidean embeddings and can cause overfitting if we attempt to solve it by adding more dimensions, Nickel & Kiela proposed using hyperbolic space instead. Hyperbolic space is characterized by constant negative curvature and is described by hyperbolic geometry. For this study, we will use the hyperbolic Poincaré disk model to embed our hierarchical data.

We develop and externally validate logistic regression and gradient boosting models across five databases: Integrated Primary Care and Information, IBM MarketScan® Medicare Supplemental, Iqvia Disease Analyzer Germany, Optum® de-identifed Clinformatics® Data Mart, and Optum® de-identified Electronic Health Record. For development, the hyperbolic embeddings are mean aggregated to be passed into the models as input. We use conditions as sole predictors, which may result in relatively low discrimination performance as compared to models using also demographic information such as age.

Results

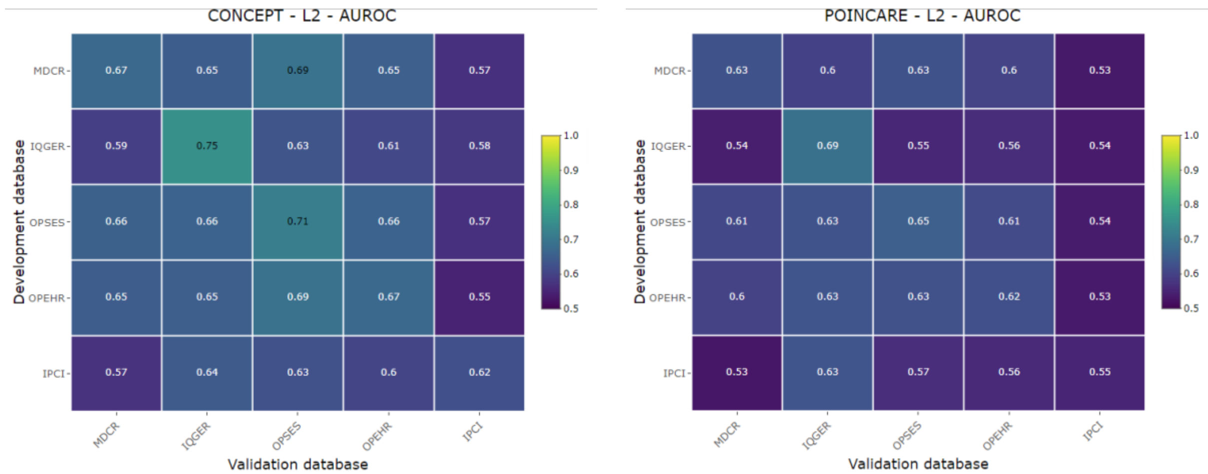

Figure 3. Discrimination of logistic regression using traditional concepts (left) and using the embeddings (right).

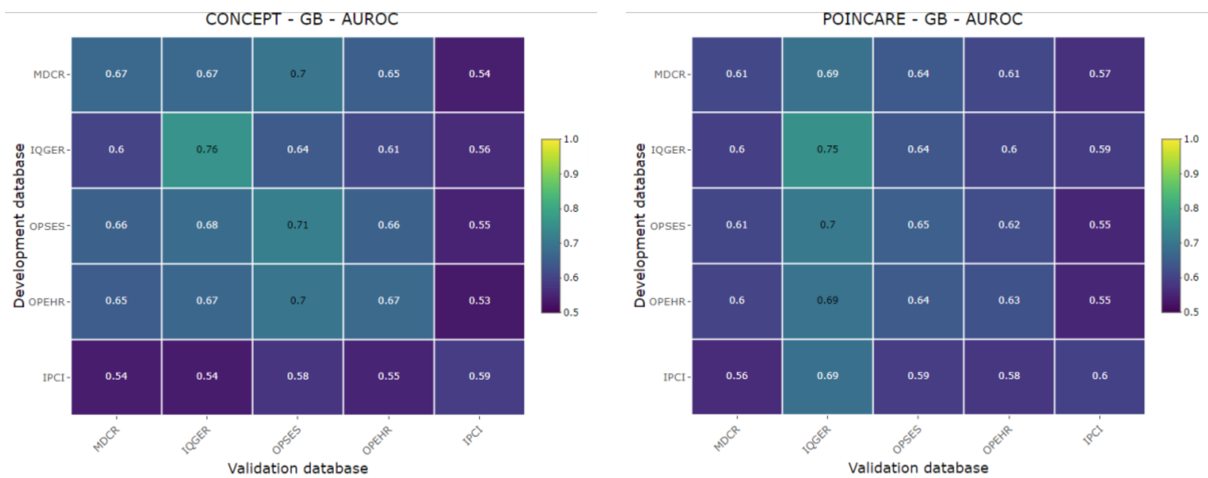

Figure 4. Discrimination of gradient boosting using traditional concepts (left) and using the embeddings (right).

Hyperbolic embeddings do not improve internal or external validation performance of logistic regression models (Figure 3). However, using gradient boosting we can observe that models trained on Integrated Primary Care and Information transport better to Iqvia Disease Analyzer Germany and vise versa. Therefore, we believe clinical domain knowledge from the SNOMED medical terms hierarchy can in some cases be used to improve external validation performance of a clinical prediction model. Future work will investigate under what exact circumstances this holds true and whether more complex models such as a Transformer will have improved validation performance, since training can be done directly on the embedding sequences. Transformers can take the embedding sequence as input directly without the mean aggregation step, which may further improve performance.

References

1. Chen, D., Liu, S., Kingsbury, P. *et al.* Deep learning and alternative learning strategies for retrospective real-world clinical data. *npj Digit. Med.* 2, 43 (2019).
2. Nickel, Maximillian, and Douwe Kiela. "Poincaré embeddings for learning hierarchical representations." *Advances in neural information processing systems* 30 (2017).