Clinical Prediction Models show Poor Generalization across a Naturalistic Sample of Outpatient Clinics.

Peter Weller[1], Kevin Hilbert[2], The KODAP Consortium, and Ulrike Lüken[1].

[1] Department of Psychology, Humboldt-Universität zu Berlin
[2] Department of Psychology, Health and Medical University Erfurt

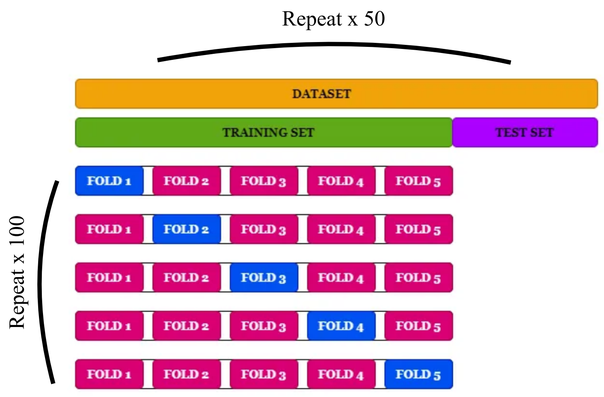
1 Background

- **Precision medicine** has the potential to enhance patient treatment efficacy. Both **patient diagnostic** and prognostic stages can be augmented with the use of model derived predictions. In a **personalized** approach to treatment selection, medical decisions are tailored based upon the outcomes of similar patient profiles that have been observed in large data sets[1].
- The extent to which models that show good performance in one clinical setting **generalize** to new patients in other contexts has been called into question. Recent studies have shown that in cross trial generalization tests, models showed poor classification accuracy despite performing above chance when tested within the same clinical trail sample that was used for model development[2,3].
- Pitfalls during model development such as **overfitting**[4] and **data leakage** between train and test can lead to biases which potentially overestimate model performance. Likewise, clinical trial datasets often represent an inaccurate reflection of patient population characteristics, due to their **stringent inclusion/exclusion criteria**.

To what extent do treatment outcome prediction models generalize across outpatient clinics in a naturalistic sample?

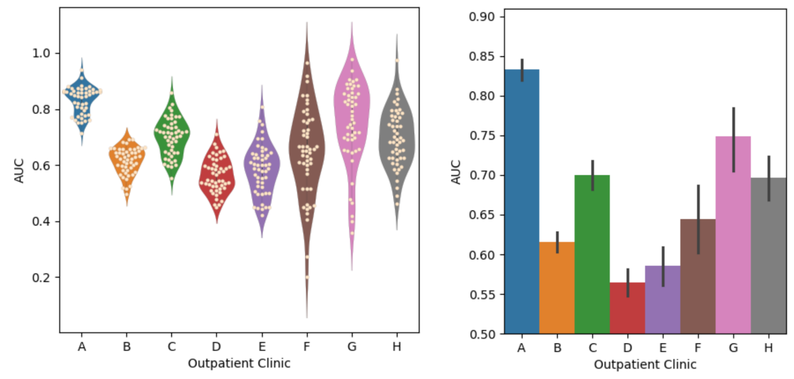
2 Methods

- A heterogeneous sample of patients with ICD-10 F3 and F4 diagnoses was collected from a selection of university outpatient clinics from numerous locations across Germany as part of the **KODAP** Network.
- We used a **machine learning** framework for model construction.
  - Training/test split.
  - Training further split into Train/ Val sets via 100x10 k-fold CV.
  - Imputation and normalization values derived from the test set.
  - **Logistic regression** with Elastic Net regularisation used.



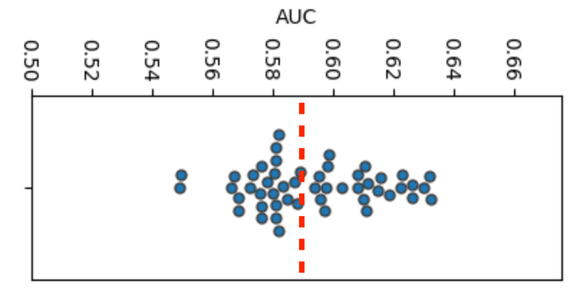
3 Results

Within-clinic generalization



We found fair to excellent within sample model generalisation across each of the clinics (mean AUC: .65, range: .56 - .83).

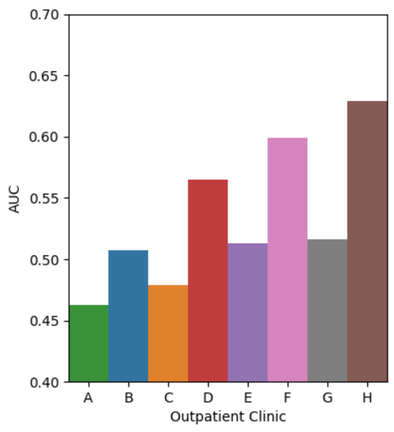
Mixed-clinic generalization



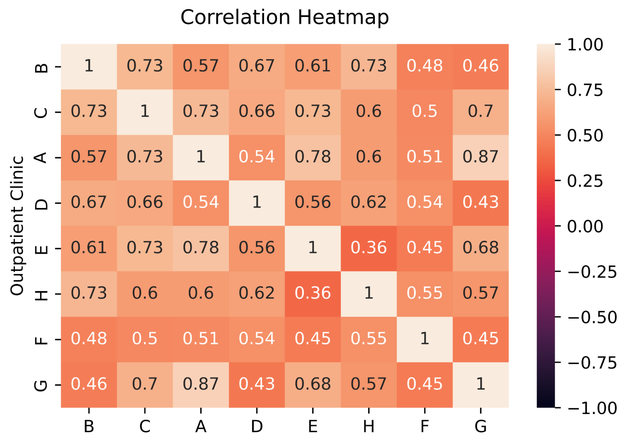
Each of the eight clinic samples combined into an aggregated dataset. models generalised above chance level, but showed a lower mean AUC (.59).

Leave-one-clinic-out generalization

Iterating across each clinic and leaving one out as a testing set, we found poor mean model generalization (mean AUC: .53. range: .46-.62).



Feature importance similarity across clinics



We calculated the **SHAP values** for each clinic and correlated **feature importance** across each of the eight clinics. Results showed both **similarities** and **differences** between pairs of clinics.

4 Discussion

- We found that prediction models reached a fair to high level of accuracy when built and tested within an individual clinical setting - replicating previous clinical prediction models built with patient data from a single clinical context[1,5]. In contrast however, whenever patient data from separate outpatient clinics was aggregated into a single sample and model generalizability across each clinic was assessed as a held out test set, model performance was poor.
- Despite the use of a more ecologically relevant dataset, a robust training, validation and testing framework, and strict control over data leakage between training and testing sets, cross clinic generalization was poor.
- These results call into question the ease with which treatment outcome prediction models can be utilised across different outpatient clinics.

References

[1] Vieira, S., Liang, X., Guiomar, R., & Mechelli, A. (2022). Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clinical psychology review*, *97*, 102193.

[2] Beliveau, V., Hedeboe, E., Fisher, P. M., Dam, V. H., Jørgensen, M. B., Frokjaer, V. G., Knudsen, G. M., & Ganz, M. (2022). Generalizability of treatment outcome prediction in major depressive disorder using structural MRI: A NeuroPharm study. NeuroImage. Clinical, 36, 103224

[3] Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science*, 383, 164–167.

[4] Lalousis, P. A., Oliver, D., Coutts, F., & Mena, S., Perfalk, E., Osimo, E. F., Daniel, W., & Perry, B. (2024). Illusory Predictions: A Case of Overfitting and Overgeneralising - Addressing Limitations from Chekroud et al. (2024). Science. 383. 164-167. 10.1126/science.adg8538#elettersSection.

[5] Hilbert, K., Kunas, S. L., Lueken, U., Kathmann, N., Fydrich, T., & Fehm, L. (2020). Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: A machine learning approach. *Behaviour research and therapy*, *124*, 103530.

The authors declare no conflict of interest.

Contact: wellerpe@hu-berlin.de