The Journal of Open Source Software

biodiscvr: Biomarker Discovery Using Composite Value Ratios

Isaac Llorente-Saguer [1] **and Neil Oxtoby** [1¶]

1 UCL Hawkes Institute and Department of Computer Science, University College London, United Kingdom ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Julia Romanowska

Reviewers:

- @donishadsmith
- @priyankagagneja

Submitted: 28 May 2025
Published: unpublished

Summary

biodiscvr provides a framework for discovering and evaluating optimised biomarkers defined as ratios of composite values derived from feature sets (e.g., regional measurements from imaging data). It was originally developed to analyse longitudinal single- and multi-cohort datasets.

The core functionality utilizes a Genetic Algorithm (GA) to search the feature space for optimal numerator and denominator combinations based on biomarker performance metrics calculated using linear mixed-effects models (Group Separation, Sample Size Estimates).

The framework allows to define inclusion criteria (e.g., in config.yaml), preprocess data, run the discovery framework, perform regional ablation analysis, and evaluate lists of biomarkers (e.g., the discovered ones) and evaluate them in multiple datasets.

Statement of need

A **composite value ratio (CVR)** (Saguer et al., 2022) is defined as the ratio between two composite aggregation of features. This concept is particularly powerful in neuroimaging, where shared confounding factors across features can cancel out, revealing more robust biomarkers. However, brute-force exploration of CVRs is computationally infeasible due to the combinatorial explosion of possible feature groupings. Encoding this search space and applying an exploratory algorithm enables the discovery of high-performing biomarkers tailored to specific clinical or research goals.

While the core methodology has been previously described, there is currently no openly available framework that implements it in a modular and extensible way. This package fills that gap by providing a flexible implementation that allows users to swap out discovery algorithms, redefine fitness metrics, and adapt the pipeline to different domains.

A key innovation in this framework is the integration of multicohort regularisation, which enhances generalisability by allowing the algorithm to be informed by multiple independent datasets without requiring data merging. This approach preserves cohort-specific structure while leveraging shared signal, making it especially valuable in heterogeneous biomedical contexts.

The current implementation relies on the following R packages: (GA (Scrucca, 2016), lme4 (Bates et al., 2015), lmmpower (Iddi & Donohue, 2022)).

Mathematics

The search algorithm is guided by the sample size estimate of a hypothetical clinical trial, and by a truncated measure of group separation, so as to avoid a Pareto frontier when trying to

optimise multiple metrics.

When using multiple cohorts for biomarker discovery, the Pareto front is dealt with using a reference direction, and multiplying the fitness function by the square of the similarity cosine with respect to the direction defined by the fitness of multiple cohorts, thus modifying the search space for convergence towards the desired equilibrium. This reference direction can be (ideally) the single best performance per cohort (which the framework can evaluate), or when none is provided, it defaults to a vector of ones (equal cohort weight).

The metrics are described in (Llorente-Saguer & Oxtoby, 2024). A linear mixed-effects model is fit to the log-transformed biomarker, and then the following metrics are assessed:

- Sample size estimate for a hypothetical clinical trial, with their parameters stated in the config.yaml file
- Group separation: it is the t-statistic of the fixed effects of being amyloid-positive
- Percentage error: standard deviation of the model residuals, as a proxy for the coefficient of variation of the biomarker in its native space.

Citations

This package builds upon the methodologies described in (Llorente-Saguer & Oxtoby, 2024).

Acknowledgements

Thank you, David Pérez Suárez, for testing the package and providing feedback. We acknowledge funding from a UKRI Future Leaders Fellowship (MR/S03546X/1, MR/X024288/1).

References

The current implementation relies on the following R packages: (GA (Scrucca, 2016), lme4 (Bates et al., 2015), lmmpower (Iddi & Donohue, 2022)).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Iddi, S., & Donohue, M. C. (2022). Power and sample size for longitudinal models in r-the longpower package and shiny app. *R Journal*, *14*(1), 264–281. https://doi.org/10.32614/RJ-2022-022

Llorente-Saguer, I., & Oxtoby, N. P. (2024). A data-driven framework for biomarker discovery applied to optimizing modern clinical and preclinical trials on alzheimer's disease. *Brain Communications*, *6*(6), fcae438. https://doi.org/10.1093/braincomms/fcae438

Saguer, I. L., Busche, M. A., & Oxtoby, N. P. (2022). Composite SUVR: A new method for boosting alzheimer's disease monitoring and diagnostic performance, applied to tau PET. *Alzheimer's & Dementia*, *18*, e063177. https://doi.org/10.1002/alz.063177

Scrucca, L. (2016). On some extensions to GA package: Hybrid optimisation, parallelisation and islands evolution. *arXiv Preprint arXiv:1605.01931*. https://doi.org/10.48550/arXiv.1605.01931