The Journal of Open Source Software

biodiscvr: Biomarker Discovery Using Composite Value Ratios

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

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

DOI: 10.21105/joss.09070

Software

- Review
- Repository
- Archive

Editor: Julia Romanowska

Reviewers:

- @donishadsmith
- @erik-whiting

Submitted: 28 May 2025
Published: 15 March 2026

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). The conceptual approach is domain-agnostic, but the current implementation is tailored to Alzheimer's disease (AD) research, reflecting the motivating use cases that guided its development. As such, several preprocessing utilities, expected variable names, and default assumptions correspond to common AD datasets (e.g., cognitive status groupings, amyloid/tau measures, and region-of-interest conventions).

The core functionality utilises 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 users to define inclusion criteria (in `config.yaml` or overriding workflow functions), preprocess data, run the discovery workflow, perform regional ablation analyses, and evaluate lists of biomarkers (including those discovered by the algorithm) across multiple datasets. Although the current release is AD-focused, the architecture is designed to support future extensions toward more domain-agnostic workflows.

Statement of need

A **composite value ratio (CVR)** (Saguer et al., 2022) is defined as the ratio between two composite aggregations of features. This concept is widely used in neuroimaging, where shared confounding factors (such as global signal intensity or individual variations in tracer pharmacokinetics) can partially cancel out when using ratios, thereby revealing more robust biological signals. Traditionally, these ratios (e.g., the Standardised Uptake Value ratio in PET) are determined based on expert knowledge and manual selection of target and reference regions.

While standard feature selection frameworks (e.g., those implementing Lasso or Elastic Net regularisation) are effective for identifying individual predictors in linear models, they are not designed to explore the combinatorial and symbolic space of feature groupings and ratios. Exhaustive search of all possible CVR combinations is computationally unfeasible; for a typical neuroimaging template with 100 regions (or features), the search space for possible ratios exceeds 10^{46} combinations. `biodiscvr` addresses this by providing an automated exploratory framework to discover high-performing CVRs tailored to specific clinical or research objectives, such as maximising group separation or minimising sample size requirements (Llorente-Saguer & Oxtoby, 2024, 2026; Saguer et al., 2022).

A novel contribution of this implementation is the introduction of **multi-cohort regularisation** (see *Methods*). Traditionally, biomarker discovery is performed on single datasets or pooled

data. Pooling often requires intensive data harmonisation to mitigate batch effects and can lead to models that overfit to the most prevalent data structure. `biodiscvr` advances the CVR framework by treating independent datasets as separate entities during optimisation. By calculating a composite fitness (objective metric) across these datasets without requiring data merging, the framework prioritises biomarkers that perform consistently across diverse populations. This approach ensures that the resulting biomarkers are not only statistically significant but also robust and generalisable.

The current implementation of biodiscvr leverages several established R packages to manage its core components. The GA package (Scrucca, 2016) serves as the default exploratory engine. While the genetic algorithm provides an efficient means of navigating the high-dimensional feature space, the framework is designed to be modular, allowing for the future integration of alternative optimisation strategies. Statistical evaluation is conducted using lme4 (Bates et al., 2015), which fits linear mixed-effects models to assess the biomarkers, while lmmpower (Iddi & Donohue, 2022) calculates the sample size requirements for hypothetical clinical trials based on these models.

Software Architecture and Workflow

`biodiscvr` is implemented as a modular R pipeline designed to handle the logistical and computational complexities of multi-cohort biomarker discovery. The framework is structured into four functional layers that manage the transition from raw data to optimised, parsimonious biomarkers:

1. Configuration and Parameter Control

The workflow is entirely parameter-driven via a central `config.yaml` file. This file defines the search space (feature sets), statistical parameters (power, effect size), and Genetic Algorithm (GA) constraints. This decoupling of logic from parameters ensures that experiments are reproducible and allows the same discovery engine to be applied to different disease domains by simply modifying the configuration schema.

2. Data Harmonisation Layer

To facilitate multi-cohort analysis without data merging, the package provides utilities for automated data preparation. The `load_datasets()` and `preprocess_data()` functions use dictionary-based matching to resolve naming inconsistencies across independent sites (e.g., mapping heterogeneous regional labels to a standard template). This layer ensures that the discovery engine operates on harmonised data structures while preserving the unique statistical properties of each original cohort.

3. The Discovery Engine

The core discovery process is managed by `run_experiments()`, which manages single-cohort or multi-cohort optimisation runs. This function interfaces with the `GA` package to navigate the high-dimensional feature space. In each generation of the search, candidate biomarkers are passed to an internal evaluation loop that fits linear mixed-effects models (via `lme4`) and calculates clinical utility metrics (via `lmmpower`). The multi-cohort consensus logic (see *Methods*) is then applied to aggregate these metrics into a single fitness score.

4. Evaluation and Parsimony Refinement

Beyond the initial discovery, the framework includes tools for post-hoc validation and refinement. The `run_ablation()` function performs a systematic "feature removal" analysis on discovered biomarkers. This process identifies the marginal contribution of each region to the overall fitness, allowing the user to simplify complex feature sets into more parsimonious biomarkers

without a significant loss in statistical power. Finally, `evaluate_biomarkers()` allows for the standardised testing of these results across independent "hold-out" datasets to confirm generalisability.

Methods

The core of `biodiscvr` is a heuristic search for an optimal feature set (biomarker) that minimises an objective function across K independent cohorts. The framework treats biomarker discovery as a multi-objective optimisation problem and transforms it into a single-objective search via a consensus fitness function.

Individual Cohort Metrics

For each cohort i, the framework fits a linear mixed-effects model to the log-transformed candidate biomarker (Saguer et al., 2022). The discovery process is guided by two primary statistical metrics:

1. **Sample Size (N):** The estimated N required for a hypothetical clinical trial to detect a predefined change. Parameters (e.g., 80% power, 20% effect size) are defined in the `config.yaml` file.
2. **Group Separation (t):** The t-statistic associated with the fixed effect of a binary group (e.g., amyloid status).

These are combined into a scalar fitness value $f_i = t/N$. To avoid a state where the algorithm cannot improve sample size without sacrificing group separation (Pareto frontier), the t-statistic is truncated at a saturation point (defaulting to 2.6, or $p \approx 0.005$). This allows the optimiser to prioritise the reduction of N once a sufficient threshold of statistical significance is reached.

Multi-cohort Regularisation

To identify biomarkers that generalise across populations without merging datasets (thus avoiding batch effects), `biodiscvr` treats the fitness scores from K cohorts as a vector $\mathbf{F} = [f_1, f_2, \dots, f_K]$. The global fitness G is calculated using the product of absolute individual fitnesses, an angular penalty, and a directional sign:

$$G = s \cdot \left(\prod_{i=1}^{K} |f_i| \right) \cdot \cos^2(\theta)$$

Where θ (Llorente-Saguer et al., 2025) is the angle between the fitness vector $\mathbf{F}$ and a **reference direction vector** $\mathbf{R}$, and s represents the directional consistency.

The following elements capture the intuition behind the global fitness function and how it enforces multi-cohort consistency:

- **Product-based Aggregation:** By using the product of fitnesses rather than the sum, the framework ensures that a biomarker must perform well across *all* cohorts. A low performance in any single cohort will heavily penalise the global score.
- **Reference Direction ($\mathbf{R}$):** This vector represents the desired balance of performance across cohorts. While it can be a vector of ones $[1, 1, \dots, 1]$, the framework defaults to the single best performance achieved per cohort, defining an adaptable performance ceiling.
- **Angular Penalty ($\cos^2 \theta$):** This term penalises candidate biomarkers with high angular deviation from the reference, effectively enforcing consensus across cohorts (i.e., avoiding overfitting to a subset of cohorts).

- **Directional Sign (s):** To ensure biological validity, s is set to 1 only if all individual cohort fitnesses f_i are positive. If any $f_i < 0$ (indicating a reversed effect direction), s is set to -1. Otherwise, the product of negative individual fitnesses could cancel out the sign.

Optimisation Strategy

The framework navigates the combinatorial space of feature ratios and groupings using an exploratory search algorithm, currently implemented via a Genetic Algorithm. The use of the directional sign s creates a "hard barrier" in the search space, guiding the algorithm away from invalid solutions. This approach enables the discovery of complex, multi-feature biomarkers that are robust across heterogeneous datasets and would be computationally intractable to identify via exhaustive search.

Prior Work

This package builds upon the methodologies described in (Llorente-Saguer & Oxtoby, 2024), and later expanded on (Llorente-Saguer & Oxtoby, 2026). The multi-cohort analysis has been possible thanks to theta (Llorente-Saguer et al., 2025), a summary metric involving multiple dimensions.

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., Arber, C., & Oxtoby, N. P. (2025). Theta, a multidimensional ratio biomarker applied to five amyloid beta peptides for investigations in familial Alzheimer's disease. *medRxiv*, 2025–2008. https://doi.org/10.1101/2025.08.06.25333131

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

Llorente-Saguer, I., & Oxtoby, N. P. (2026). Enhanced monitoring of Alzheimer's disease brain atrophy using composite value ratios of volumes. *Brain Communications*, *8*(1), fcaf497. https://doi.org/10.1093/braincomms/fcaf497

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