Carl von Ossietzky Universität Oldenburg

DEPARTMENT OF PSYCHOLOGY
CARL VON OSSIETZKY UNIVERSITY OLDENBURG

Multiverse Analysis: Outcome Visualization and Applications in Neuroimaging

Cassie Short, Daniel Kristanto, Micha Burkhardt, Andrea Hildebrandt

17-18.06.2025

The schedule for the afternoon

- 13:30-15:00: Multiverse Analysis Theory and Conceptualisation
- 15:00-15:30: Coffee break
- 15:30-17:00: Multiverse Analysis in Neuroimaging
- 17:00-18:00: Conceptualisation activity

Content

Key Terminology

In a multiverse analysis, we consider different analytical decisions.

- **Decision Node:** Refers to any decision point in the analysis pipeline where multiple options are possible (e.g., handling outliers).
- **Option:** Refers to the specific setting or choice selected for a given decision node (e.g., "remove outliers" vs. "keep outliers").
- A complete analysis, or pipeline, or "universe," is defined by one unique combination of options across all decision nodes.

Key Tasks in Multiverse Analysis

The goals of a multiverse analysis can be grouped into several key tasks:

1. **Understand Composition:** Identify the different analysis **decision nodes** and the **options** for each.
2. **Assess Outcome Sensitivity:** See how much the results vary. Are the conclusions robust?
3. **Connect to Outcomes:** Identify which **options** are the main drivers of variation in the results.
4. **Connect Combinations:** Investigate if interactions between **options** cause specific outcomes.
5. **Validate the Multiverse:** Evaluate the reasonableness of the different universes.

Visualization 1: Outcome Distribution

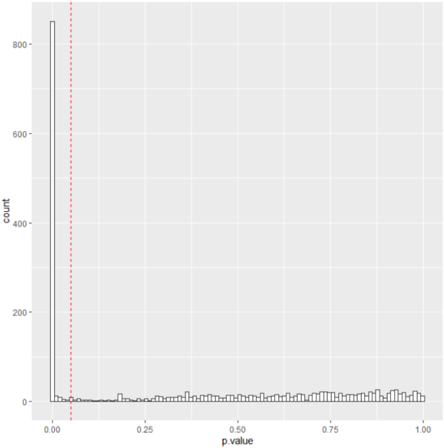


Figure: An Outcome Histogram shows the frequency of different outcomes across the multiverse.

Outcome Distribution: Explanation

Explanation:

- This visualization shows the distribution of a specific outcome (such as p-values) from all analyses in the multiverse.
- It is often shown as a histogram or a density plot.

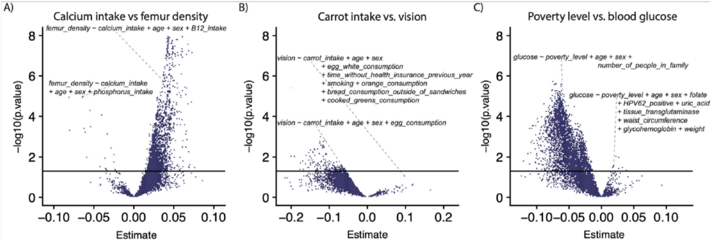
Key Features:

- Quickly shows the range of possible outcomes.
- Highlights the most common results.

Why this is not enough?

- A distribution of p-values doesn't show the **magnitude** of the effects. A result can be significant but trivial.
- It doesn't tell us **which options** for each decision node lead to these outcomes (**connect** task).

Visualization 2: Vibration of Effects Plot



Tierney et al. (2021)

Figure: A Vibration of Effects Plot shows the relationship between effect size and statistical significance.

Vibration of Effects Plot: Explanation

Explanation:

- A scatter plot that shows the relationship between two outcomes simultaneously, typically effect size (x-axis) and statistical significance (y-axis).
- Each point represents one combination of options.

Key Features:

- Shows the "vibration" of results; how effect size and significance co-vary.
- Helps identify if stronger effects are also more significant.

Why this is not enough?

- While we can see how outcomes cluster, it's hard to know the exact **combination of options** that produced a specific result.
- We need a way to see the full "recipe" for each outcome.

Visualization 3: Specification Curve

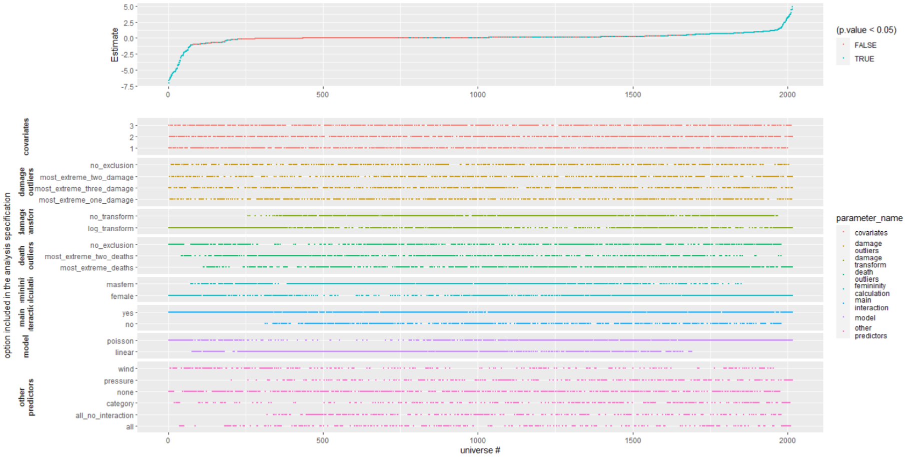


Figure: A Specification Curve connects each outcome to its specific set of options.

Specification Curve: Explanation

Explanation:

- A composite visualization that connects each specific combination of **options** directly to their outcomes.
- It has two main linked panels: an Outcome Curve (top) and a Specification Panel (bottom).

Key Features:

- Directly links **outcomes to specific options**.
- Excellent for identifying which options (for any given decision node) drive the results.
- Shows the full range of outcomes while also revealing the underlying recipe (the set of options) for each one.

Specification Curve: Challenges

Distinction between probabilistic and possibilistic interpretations of multiverse analysis results

- **Probabilistic:** Viewing outcomes as having certain probabilities based on their frequencies within the set of explored universes. Akin to statistical likelihood.
  - **Probabilistic interpretation:** Assumes all explored universes (or analytical specifications) in a multiverse analysis are equally likely. If an outcome appears more frequently among these universes, it might be mistakenly interpreted as being more likely correct.
- **Possibilistic:** Viewing outcomes as possibilities that arise from different reasonable analytical decisions. It does not assign likelihood but rather acknowledges the potential validity of various outcomes without preference.
  - **Possibilistic interpretation:** Suggests that an outcome's presence simply indicates it is a possible result of reasonable analytical choices, without assigning likelihood based on frequency.

Specification Curve: Challenges

- These more traditional visualisations may express data in a way that encourages probabilistic interpretations, even when such an interpretation is inappropriate for multiverse data (Hall et al., 2022).
- **Illusion of Probability:** The potential misunderstanding that might arise **when frequency information in visualizations is interpreted as indicating the likelihood of correctness**, which is not a valid assumption in multiverse analysis due to the non-random and interdependent nature of the analyzed universes.
- "a principled interpretation of the multiverse analysis results considers **the variation in outcomes as possibilistic, and the uncertainty in each individual outcome as probabilistic**" (Sarma et al., 2024)

Hall et al. (2022) A survey of tasks and visualizations in multiverse analysis reports. *Computer Graphics Forum*, 41(1), 402-426.

Specification Curve: Challenges

How do we visualise a multiverse of results that conveys the valuable information about outcome frequency without misleading viewers into probabilistic interpretations?
How do we include information that is important for other tasks in multiverse analysis (connect, validate, interpret) in a multiverse analysis report?

Miliways Multiverse Visualization

Connect, Validate, Interpret: Milliways Package
The Milliways Package (Sarma et al., 2024) visualises a multiverse analysis based on two principles:

Readers of a multiverse analysis should be able to

1. assess whether the decisions in the analysis are all equally justifiable
2. correctly distinguish between the probabilistic and possibilistic uncertainty inherent in such an analysis

The Milliways Package provides a visualisation design that aims to do this.

Sarma et al. (2024) Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA.

Miliways Multiverse Visualization

To visualise the results of a multiverse analysis using Milliways, users would need to provide as input the results of a multiverse analysis, the analysis code, and the dataset used for the analysis.

The results file should contain, for each universe:

1. the option name for every parameter
2. the mean point estimate
3. x and y values for the consonance curve

Sarma et al. (2024) Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA.

Milliways Package

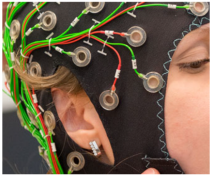
Let's take a look at their template!

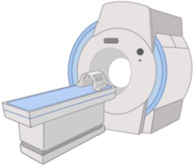
You can download it for yourself here:
https://osf.io/y2cmt?view_only=bd4668d3241c43e4b699dd4e1f88477a

Or you can follow the demo on the screen

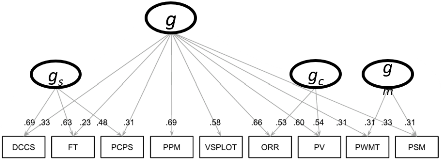
Sarma et al. (2024) Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA.

Cognitive Neuroimaging: Goal



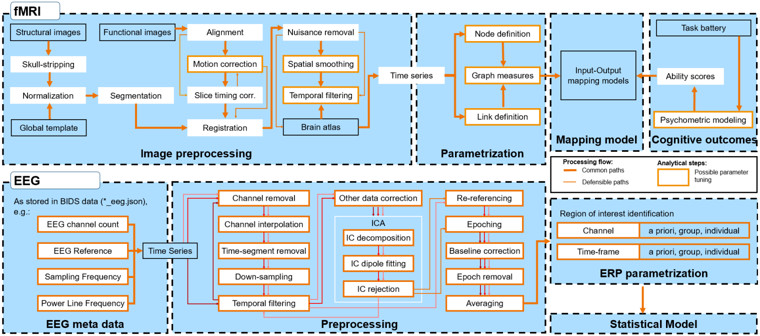


Neuroimaging Features



Cognitive Neuroimaging: Challenges

- The need of preprocessing pipeline to remove the unwanted signals.
- There are many options to preprocess neuroimaging data
- Different preprocessing pipelines may lead to different results



Example 1: Botvinik-Nezer et al. (2020)

Article

Variability in the analysis of a single neuroimaging dataset by many teams

https://doi.org/10.1038/s41586-020-2314-9

Received: 14 November 2019

Accepted: 7 April 2020

Published online: 20 May 2020

Check for updates

A list of authors and affiliations appears in the online version of the paper.

Data analysis workflows in many scientific domains have become increasingly complex and flexible. Here we assess the effect of this flexibility on the results of functional magnetic resonance imaging by asking 70 independent teams to analyse the same dataset, testing the same 9 ex-ante hypotheses[1]. The flexibility of analytical approaches is exemplified by the fact that no two teams chose identical workflows to analyse the data. This flexibility resulted in sizeable variation in the results of hypothesis tests, even for teams whose statistical maps were highly correlated at intermediate stages of the analysis pipeline. Variation in reported results was related to several aspects of analysis methodology. Notably, a meta-analytical approach that aggregated information across teams yielded a significant consensus in activated regions. Furthermore, prediction markets of researchers in the field revealed an overestimation of the likelihood of significant findings, even by researchers with direct knowledge of the dataset[2–5]. Our findings show that analytical flexibility can have substantial effects on scientific conclusions, and identify factors that may be related to variability in the analysis of functional magnetic resonance imaging. The results emphasize the importance of validating and sharing complex analysis workflows, and demonstrate the need for performing and reporting multiple analyses of the same data. Potential approaches that could be used to mitigate issues related to analytical variability are discussed.

Example 1: Botvinik-Nezer et al. (2020)

- Same dataset independently analyzed by 70 teams to test 9 hypotheses about brain activity in a risky-decision task
- High variability with respect to the reported statistically significant result

Hypotheses	# of teams	% of teams
1	26	37.1
2	15	21.4
3	16	22.9
4	23	32.9
5	59	84.3
6	23	32.9
7	4	5.7
8	4	5.7
9	4	5.7

Example 1: Botvinik-Nezer et al. (2020)

Hypotheses	# of teams	% of teams
1	26	37.1
2	15	21.4
3	16	22.9
4	23	32.9
5	59	84.3
6	23	32.9
7	4	5.7
8	4	5.7
9	4	5.7

- About 20% of the analyses came to a conclusion opposite to that of the majority
- Three most ambiguous hypotheses highlighted

Example 2: Trübutschek et al. (2025)

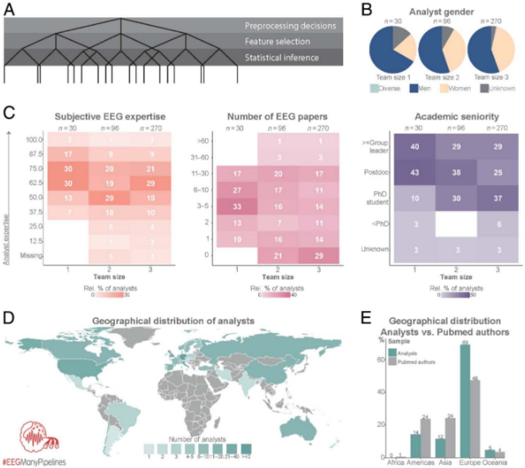
EEGManyPipelines: A Large-scale, Grassroots Multi-analyst Study of Electroencephalography Analysis Practices in the Wild

Darinka Trübutschek[1*], Yu-Fang Yang[2*], Claudia Gianelli[3*], Elena Cesnaite[4], Nastassja L. Fischer[5], Mikkel C. Vinding[6,7], Tom R. Marshall[8], Johannes Algermissen[9,10], Annalisa Pascarella[11], Tuomas Puoliväli[12], Andrea Vitale[13], Niko A. Busch[4†], and Gustav Nilsonne[7,14†]

Abstract

■ The ongoing reproducibility crisis in psychology and cognitive neuroscience has sparked increasing calls to re-evaluate and reshape scientific culture and practices. Heeding those calls, we have recently launched the EEGManyPipelines project as a means to assess the robustness of EEG research in naturalistic conditions and experiment with an alternative model of conducting scientific research. One hundred sixty-eight analyst teams, encompassing 396 individual researchers from 37 countries, independently analyzed the same unpublished, representative EEG data set to test the same set of predefined hypotheses and then provided their analysis pipelines and reported outcomes. Here, we lay out how large-scale scientific projects can be set up in a grassroots, community-driven manner without a central organizing laboratory. We explain our recruitment strategy, our guidance for analysts, the eventual outputs of this project, and how it might have a lasting impact on the field. ■

Example 2: Trübutschek et al. (2025)



The Need for Multiverse Analysis in Neuroimaging

- Given the high variability of results depending on the chosen analysis pipeline, the implementation of **multiverse analysis** in neuroimaging is growing.
- The first step of performing a multiverse analysis is to define the "multiverse" itself.
- This means creating a comprehensive list of all reasonable and available data processing and analysis pipelines that could be applied.

The Multiverse in Graph-based fMRI

- Systematic literature review of graph theory-based functional Magnetic Resonance Imaging (fMRI) studies.
- A total of **252** studies coded in terms of their analytical pipeline.
- The multiplicity of analytic pipelines summarized in a Shiny app.

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

ELSEVIER

The multiverse of data preprocessing and analysis in graph-based fMRI: A systematic literature review of analytical choices fed into a decision support tool for informed analysis

Daniel Kristanto[a,*], Micha Burkhardt[a], Christiane Thiel[a,b,c], Stefan Debener[a,b,c], Carsten Gießing[a,b,1], Andrea Hildebrandt[a,b,c,1]

[a] Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg 26129, Germany
[b] Research Center Neurosensory Science, Carl von Ossietzky Universität Oldenburg, Germany
[c] Cluster of Excellence "Hearing4All", Carl von Ossietzky Universität Oldenburg, Germany

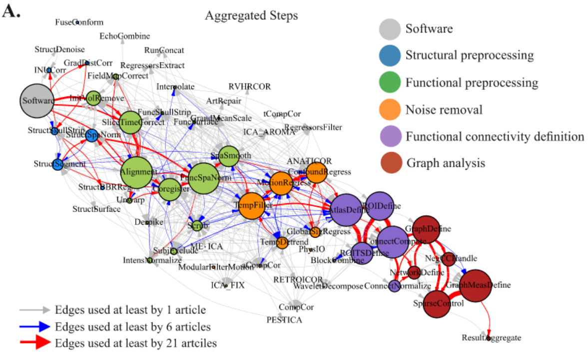
ARTICLE INFO

Keywords:
Functional magnetic resonance imaging – fMRI
Network neuroscience
Graph theory
Robustness
Multiverse analysis
Forking paths
Data preprocessing
Shiny app

ABSTRACT

The large number of different analytical choices used by researchers is partly responsible for the challenge of replication in neuroimaging studies. For an exhaustive robustness analysis, knowledge of the full space of analytical options is essential. We conducted a systematic literature review to identify the analytical decisions in functional neuroimaging data preprocessing and analysis in the emerging field of cognitive network neuroscience. We found 61 different steps, with 17 of them having debatable parameter choices. Scrubbing, global signal regression, and spatial smoothing are among the controversial steps. There is no standardized order in which different steps are applied, and the parameter settings within several steps vary widely across studies. By aggregating the pipelines across studies, we propose three taxonomic levels to categorize analytical choices: 1) inclusion or exclusion of specific steps, 2) parameter tuning within steps, and 3) distinct sequencing of steps. We have developed a decision support application with high educational value called METEOR to facilitate access to the data in order to design well-informed robustness (multiverse) analysis.

Guided Exploration



METEOR Shiny App

https://www.apps.meta-rep.lmu.de/METEOR/

The Multiverse in (Mobile) EEG

- Systematic literature review of mobile electroencephalography (EEG) studies analyzing P3 event related potential (ERP) during walking and standing.
- A total of **27** studies coded in terms of their analytical pipeline.
- The multiplicity of analytic pipelines summarized in a Shiny app.

DOI: 10.1111/psyp.14743

REVIEW

PSYCHOPHYSIOLOGY SPR WILEY

Preprocessing choices for P3 analyses with mobile EEG: A systematic literature review and interactive exploration

Nadine S. J. Jacobsen[1] | Daniel Kristanto[2] | Suong Welp[1] | Yusuf Cosku Inceler[2] | Stefan Debener[1,3,4]

[1]Neuropsychology Lab, Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
[2]Psychological Methods and Statistics Division, Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
[3]Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
[4]Centre for Neurosensory Science & Systems, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

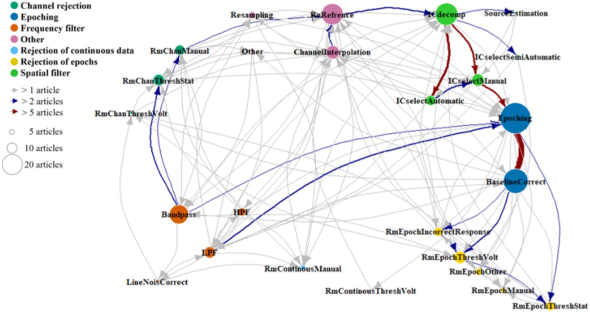
Correspondence
Nadine S. J. Jacobsen, Neuropsychology Lab, Department of Psychology, Carl von Ossietzky Universität Oldenburg, Ammerländer Heerstraße 114-118, Oldenburg 26129, Germany.
Email: nadine.jacobsen@uni-oldenburg.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: DE 779/8-1

Abstract
Preprocessing is necessary to extract meaningful results from electroencephalography (EEG) data. With many possible preprocessing choices, their impact on outcomes is fundamental. While previous studies have explored the effects of preprocessing on stationary EEG data, this research delves into mobile EEG, where complex processing is necessary to address motion artifacts. Specifically, we describe the preprocessing choices studies reported for analyzing the P3 event-related potential (ERP) during walking and standing. A systematic review of 258 studies of the P3 during walking, identified 27 studies meeting the inclusion criteria. Two independent coders extracted preprocessing choices reported in each study. Analysis of preprocessing choices revealed commonalities and differences, such as the widespread use of offline filters but limited application of line noise correction (3 of 27 studies). Notably, 59% of studies involved manual processing steps, and 56% omitted reporting critical parameters for at least one step. All studies employed unique preprocessing strategies. These findings align with stationary EEG preprocessing results, emphasizing the necessity for standardized reporting in mobile EEG research. We implemented an interactive visualization tool (Shiny app) to aid the exploration of the preprocessing landscape. The app allows users to structure the literature regarding different processing steps, enter planned processing methods, and compare them with the literature. The app could be utilized to examine how these choices impact P3 results and understand the robustness of various processing options. We hope to increase awareness regarding the potential influence of preprocessing decisions and advocate for comprehensive reporting standards to foster reproducibility in mobile EEG research.

KEYWORDS
EEG, electroencephalography, mobile, preprocessing, shiny app, systematic literature review

The Multiverse in (Mobile) EEG



Guided Exploration

METEOR-EEG Shiny App

`https://meteor-eeg-oldenburg.shinyapps.io/preprocessing/`

Multiverse Analysis in fMRI: Example

- Dafflon et al. performed multiverse analysis to predict brain age based on graph measures derived from resting-state fMRI
- A total of **544** different pipelines were implemented

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-022-31347-8 OPEN

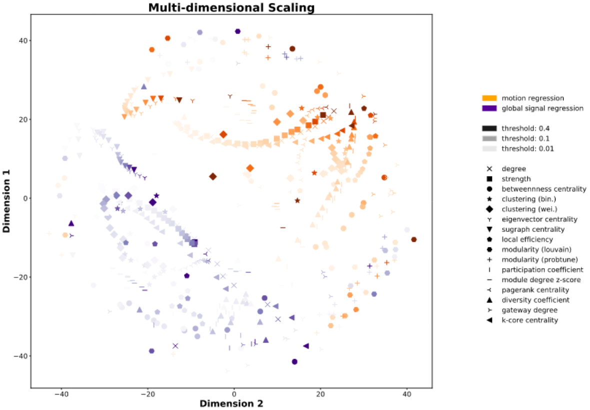
A guided multiverse study of neuroimaging analyses

Jessica Dafflon [1✉], Pedro F. Da Costa [1,2], František Váša [1], Ricardo Pio Monti[3], Danilo Bzdok [4,5], Peter J. Hellyer[1], Federico Turkheimer[1], Jonathan Smallwood [6], Emily Jones[2] & Robert Leech [1✉]

For most neuroimaging questions the range of possible analytic choices makes it unclear how to evaluate conclusions from any single analytic method. One possible way to address this issue is to evaluate all possible analyses using a multiverse approach, however, this can be computationally challenging and sequential analyses on the same data can compromise predictive power. Here, we establish how active learning on a low-dimensional space capturing the inter-relationships between pipelines can efficiently approximate the full spectrum of analyses. This approach balances the benefits of a multiverse analysis without incurring the cost on computational and predictive power. We illustrate this approach with two functional MRI datasets (predicting brain age and autism diagnosis) demonstrating how a multiverse of analyses can be efficiently navigated and mapped out using active learning. Furthermore, our presented approach not only identifies the subset of analysis techniques that are best able to predict age or classify individuals with autism spectrum disorder and healthy controls, but it also allows the relationships between analyses to be quantified.

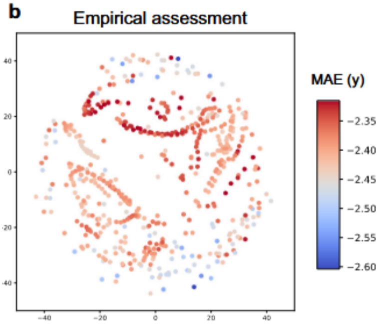
Low-Dimensional Representation of the Pipelines

- The outputs (graph measures) of each pipeline were visualized in **two-dimensional space**, where each dot corresponds to one pipeline.
- Pipelines that are close to each other in the space are more similar.



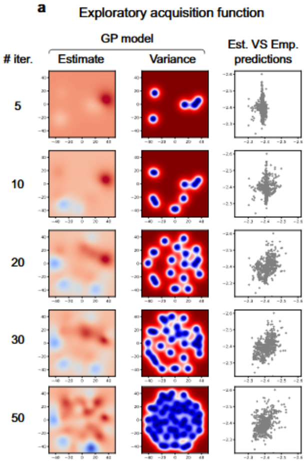
Prediction Performance

- For each pipeline, the graph measures were used to predict brain age.
- The prediction performance was evaluated by using **Mean Absolute Error (MAE)** between actual and predicted brain age.



Active Leaning for Efficient Multiverse Analysis

- Dafflon et al. also proposed **active learning algorithm** to run multiverse analysis efficiently.
- This algorithm works by **intelligently sampling** subset of the pipelines to estimate the outcome of the whole multiverse.



Multiverse Analysis in EEG: Example

- Short et al. performed multiverse analysis to predict extraversion scores from the Late Positive Potential.
- A total of **528** different pipelines were implemented.

Lost in a Large EEG Multiverse? Comparing Sampling Approaches for Representative Pipeline Selection

Cassie Ann Short, Andrea Hildebrandt, Robin Bosse, Stefan Debener, Metin Özyağcılar, Katharina Paul, Jan Wacker, Daniel Kristanto

doi: https://doi.org/10.1101/2025.04.08.647779

This article is a preprint and has not been certified by peer review [what does this mean?].

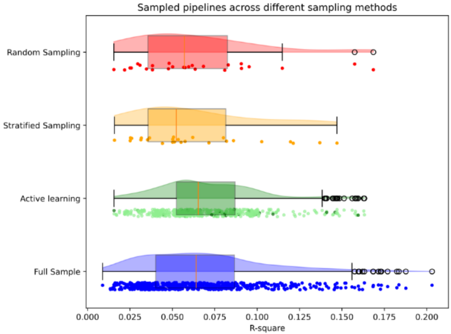
Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

The multiplicity of defensible strategies for processing and analysing data has been implicated as a core contributor to the replicability crisis, creating uncertainty about the robustness of a result to variations in data processing choices. This issue is exacerbated where a large number of data processing pipelines are defensible, and where there is great heterogeneity in the pipelines applied in the literature, such as in processing and analysing electroencephalography (EEG) signals. In a multiverse analysis, equally defensible pipelines are computed and the robustness of the result to these variations is reported. However, a large number of defensible pipelines is sometimes infeasible to compute exhaustively, and researchers rely on sampling approaches. In these cases, pipelines are sampled from the full multiverse and the robustness is reported across these samples, assuming that they are representative for the entire multiverse. However, different sampling methods may yield different robustness results, introducing what we term multiverse sampling uncertainty. To

Multiverse Analysis in EEG: Example

- The main goal of the paper is to highlight **variability in the representativeness** of the distribution of model fits between different sampling approaches in multiverse analysis: **random sampling, stratified sampling, and active learning**.
- The active learning sample most closely represented the median model fit of the full multiverse.



Summary

- There are many ways to visualize multiverse analysis outcomes. Interactive platforms such as **Miliways** allow for a deeper exploration of not just the results, but also the relationship between outcomes and the analytical decisions that produced them.

- In **neuroimaging**, multiverse analysis is a growing necessity due to the vast number of available analytical decisions and the significant result variability they cause.

- To ease the computational burden of running so many pipelines, methods such as **active learning** are being explored to make multiverse analysis more efficient and practical to implement.

Don't get lost in the Garden of Forking paths

Think carefully about the analytical choices you can make!

Thank you for attending this workshop!

andrea.hildebrandt@uol.de
cassie.ann.short@uol.de
daniel.kristanto@uol.de
micha.burkhardt@uol.de