



**HAL**  
open science

## **HDMAX2, an R package to perform high dimension mediation analysis**

Florence Pittion, Basile Jumentier, Aurélie Nakamura, Johanna Lepeule,  
Olivier François, Magali Richard

► **To cite this version:**

Florence Pittion, Basile Jumentier, Aurélie Nakamura, Johanna Lepeule, Olivier François, et al.. HDMAX2, an R package to perform high dimension mediation analysis. COBICA 2024, Mar 2024, Grenoble, France. hal-04664854

**HAL Id: hal-04664854**

**<https://hal.science/hal-04664854v1>**

Submitted on 30 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# HDMAX2, an R package to perform high dimension mediation analysis

Florence Pittion, Basile Jumentier, Aurélie Nakamura,  
Johanna Lepeule, Olivier François, Magali Richard

MAGe team, TIMC, CNRS UMR 5525, Grenoble

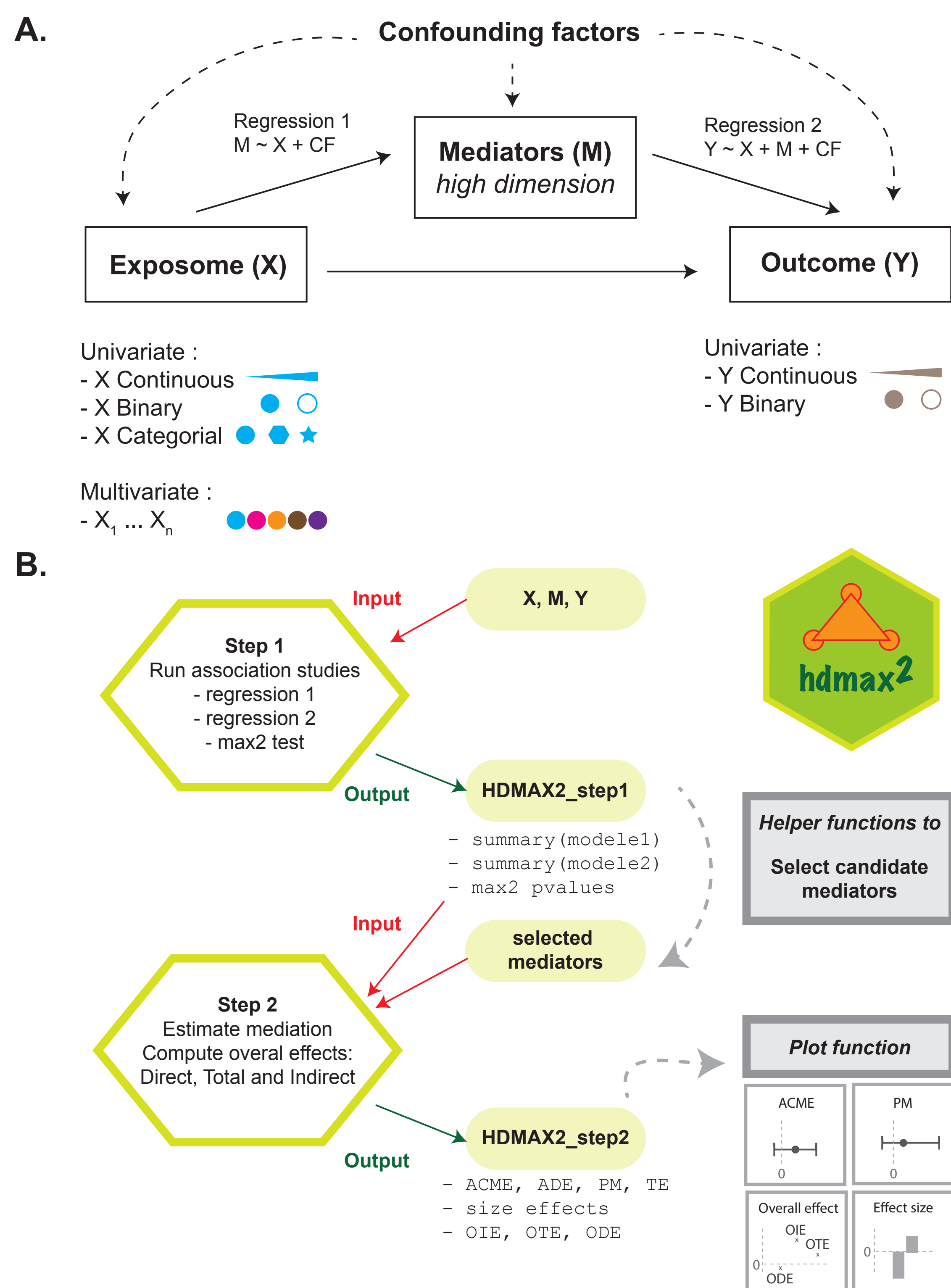
corresponding author: [florence.pittion@univ-grenoble-alpes.fr](mailto:florence.pittion@univ-grenoble-alpes.fr)



## Introduction

Mediation analysis is crucial in epidemiology for understanding how exposures impact health outcomes. There is an increasing interest in applying mediation analysis to explore the causal relationships between **patient environmental exposure**, **molecular features** (such as omics data) and patients outcomes. The **high-dimension** of omic molecular features presents statistical challenges. Despite various methods addressing these challenges, most are limited to continuous outcomes, and their adoption among epidemiologists and health data scientists is limited. To address this gap, we introduce an R package specifically tailored for high-dimensional mediation analysis, leveraging the **HDMAX2** method (Jumentier et al. 2023). The package called `hdmax2` is enhanced to handle **multivariate exposure** and **non-continuous outcomes**, making it applicable to diverse healthcare data scenarios and improving accessibility for users with varying expertise.

## HDMAX2



A. Model of the `hdmax2` R package. B. Core structure of the package.

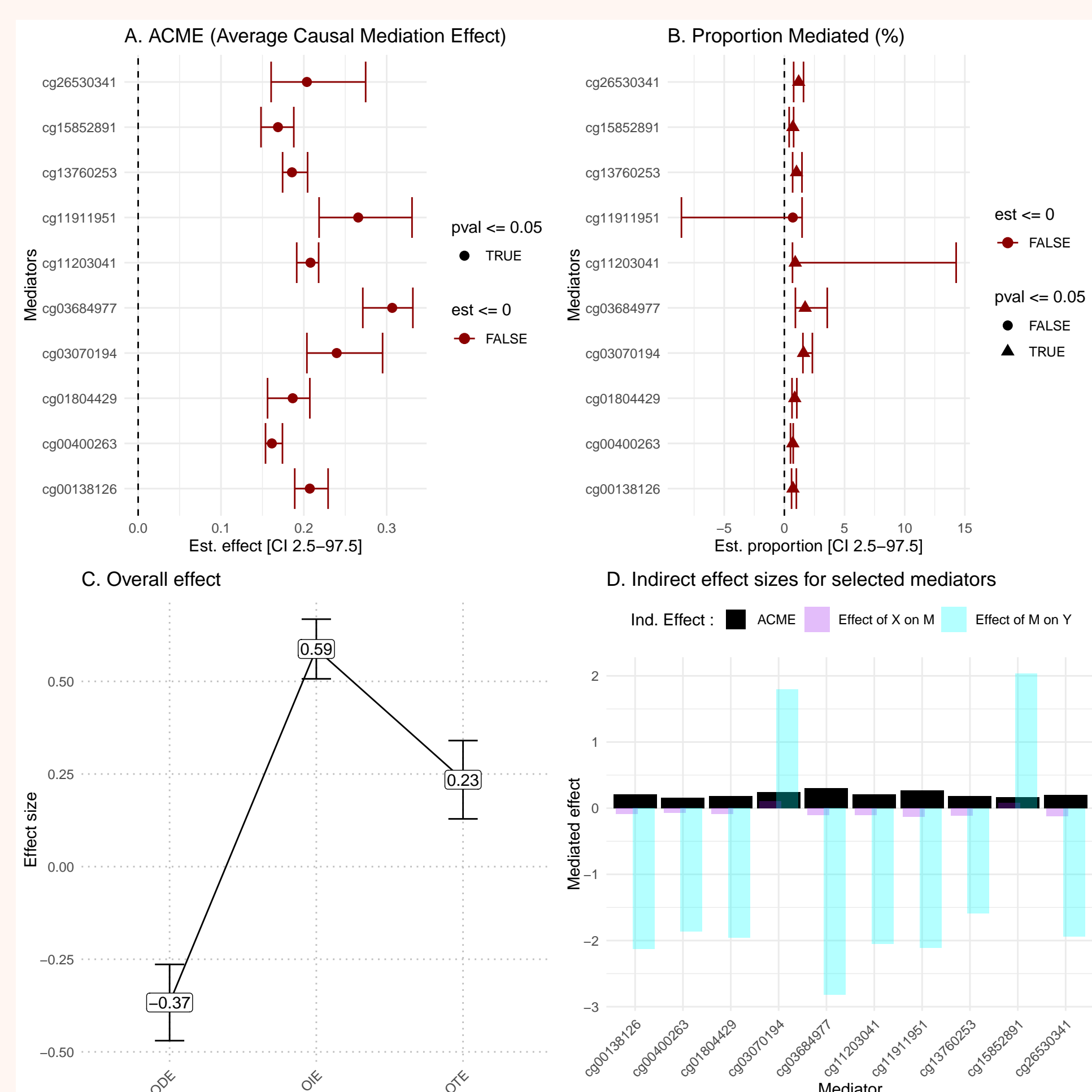
## Method

- Input data.** The HDMAX2 package is designed to accept exposure  $X$  consisting of univariate data, which can be continuous, binary, or categorical, as well as multivariate exposures. Binary variables are converted to 0s and 1s and treated as continuous variables.
- Step 1: Identification of potential mediators.** The `run_AS` function evaluates associations between exposure, mediators and outcome, using Latent Factor Mixed Models (LFMMs) (Caye et al. 2019). It performs two latent factor regression and a significance statistical test *max-squared*. The return object `HDMAX2-step1` includes the following variables:
  - $U$ , scores matrix for the  $K$  latent factors,  $V$ , their loadings
  - For each regression: GIF, Genomic Inflation Factor, p-value, z-score, and f-score and adjusted  $R^2$ .
  - The p-value of the *max-squared* mediation test.
  - Input variables of the model:  $X$  (Exposure),  $Y$  (Outcome),  $M$  (potential Mediators) and  $C$  (Covariates, when applicable).
- Step 2: Estimation of the mediator indirect effects.** The function `estimate_effect` takes into input the object `HDMAX2-step1` (output from the step1) and a list of potential mediators  $M_s$  to analyse.  $M_s$  is a subset of  $M$  that is defined by the user using `HDMAX2-step1` output. The `estimate_effect` function returns an object of type `HDMAX2-step2`, including the following variables:
  - Mediation analysis for each mediator: Average Causal Mediated Effect (ACME), Average Direct Effect (ADE), Proportion mediated (PM) and Total effect (TE),
  - The size effects of the regression models
  - Overall effects of the global mediation model: Indirect Effect (OIE), Total Effect (OTE), Direct Effect (ODE)
- Results display.** `plot_hdmax2` function takes in input an object `HDMAX2-step2` and enables graphical visualisation of mediated effects (see application below).



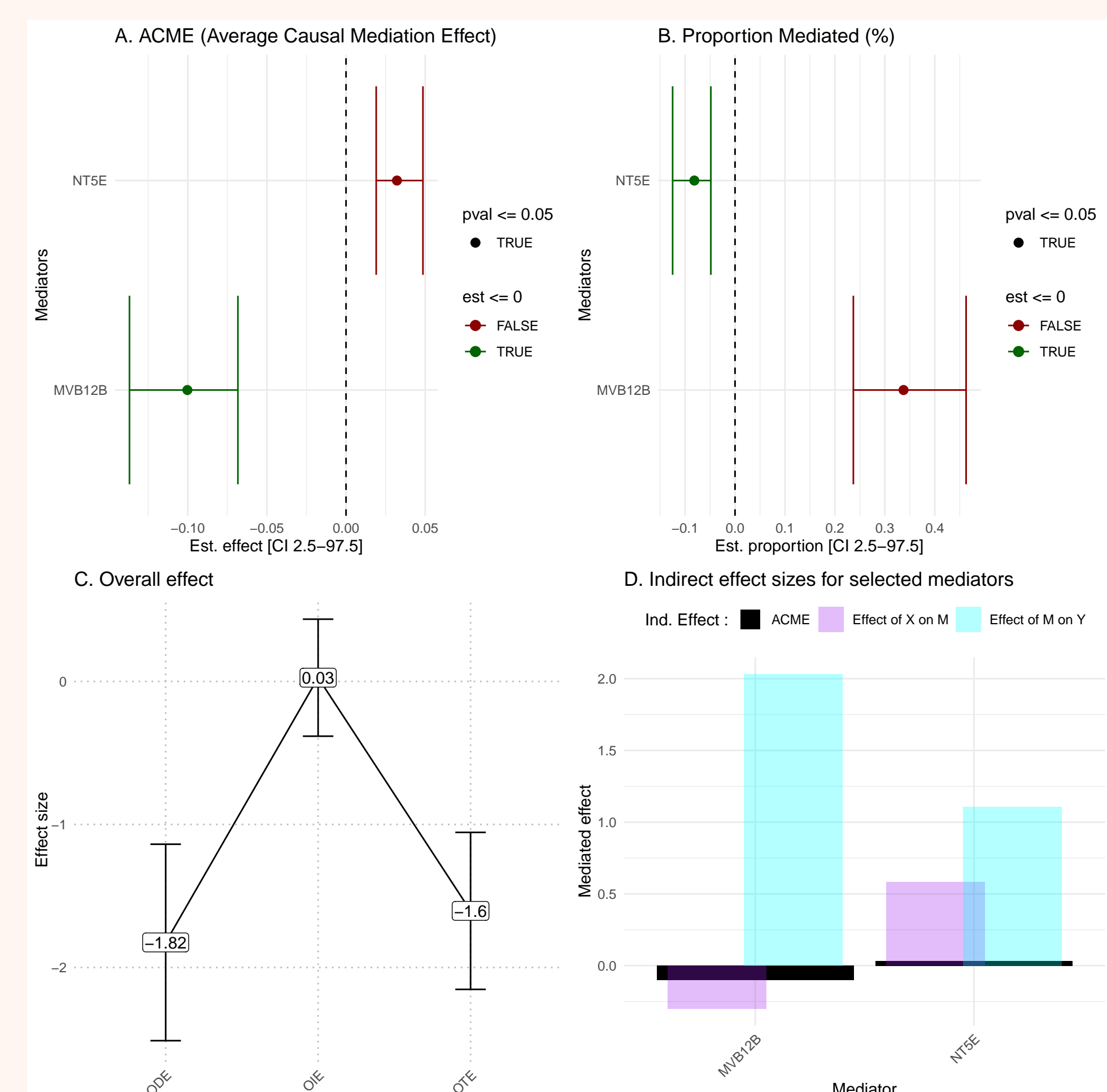
Package available: <https://github.com/bcm-uga/hdmax2>

## Breast cancer application with binary exposure, continuous outcome and DNA methylation probes as potential mediators



Mediation analysis was employed to explore the potential causal role of DNAm in connecting HER-2 (Human Epidermal Growth Factor Receptor-2) status of Breast Cancer to a survival prognostic factor, the risk score, using data from the TCGA-BRCA repository. The risk score, derived from a six-gene expression signature (Yin et al. 2024), is inversely correlated with patient survival. Categorizing patients by HER-2 status, the analysis revealed a negative direct effect of HER2 on the risk score. However, mediation by the top 10 CpG probes resulted in a detrimental effect on patient survival (increased risk), showcasing the complex interplay within biological pathways involved in tumor growth and patient survival.

## Multiple Sclerosis application with binary exposure, binary outcome and gene expression as potential mediators



Mediation analysis was conducted on gene expression (18,010 coding gene) to assess the indirect effect of gender on Multiple Sclerosis (MS) subtypes (Clinically Isolated Syndrome (CIS) and Relapsing-Remitting (RR)) (Kim et al. 2021). We focused on the top two mediators: *NT5E* and *MVB12B* which exhibit opposite effects on the pathway between gender and MS subtypes. *NT5E* is known to play a role in immune response deregulation in MS and *MVB12B* was recently identified in a genome-wide protein QTL study studying drivers of immune-related diseases. These findings could provide insightful understanding of biological mechanisms involved.