SM Journal of Environmental Toxicology

Review Article

Elja: An R Package to Perform Environment-Wide Association Studies (Ewas / Envwas) Analysis

Marwan El Homsi[1*] and Isabella Annesi-Maesano[1,2]

[1]Desbrest Institute of Epidemiology and Public Health, Univ Montpellier, INSERM, Montpellier, France

[2]Division of Respiratory Medicine, Allergology, and of Thoracic Oncology, University Hospital of Montpellier, Montpellier

Abstract

Environment-Wide Associations Studies (EWAS / EnvWAS) are an approach involving multiple analyses of the relationship between several exposures to a health event. This approach allows us to study the impact of an exposome on a health event. The Elja package was developed using R to perform the EnvWAS analytical approach. This package performs repeated analyses in the case of linear regression models, logistic regression models, and, more generally, generalized linear models.

The results are presented in two complementary ways. In a detailed way, with a table automatically generated at the end of the analysis allows all results to be consulted and then extracted. In a more visual way, with a Manhattan plot displaying the results in their entirety, with a significance threshold of 0.05 and the corrected threshold.

This package is compatible with Windows, Ubuntu and MacOS. It is available free of charge on GitHub (https://github.com/EHMarwan/Elja) and the CRAN directory.

Keywords : Epidemiology; Exposome; R; EnvWAS; EWAS; Package

Introduction

The impact of environmental factors on human health is a major public health concern. The most common studies have attempted to estimate the degree of association between an isolated environmental factor and a health event. However, this approach is insufficient for explaining multifactorial diseases. The need to consider the exposome–that is, all the exposures to which an individual is subjected over the course of his or her life–has emerged [1,2]. To achieve this, a new approach has been developed that aims to consider several environmental factors at once in relation to a health event in order to understand their overall impact on this event. The exposome includes all external factors, including chemical, physical, biological, and social elements, which can potentially influence an individual's health and well-being. In the world of exposome studies, one methodology is increasingly in use: Environmental-Wide Association Studies (EWAS, later called EnvWAS).

Submitted: 17 March 2024 | **Accepted:** 23 March 2024 | **Published:** 25 March 2024

***Corresponding author:** Marwan El Homsi, Desbrest Institute of Epidemiology and Public Health, Univ Montpellier, INSERM, Montpellier, France

Citation: El Homsi, Annesi-Maesano I.(2024)Elja: An R Package to Perform Environment-Wide Association Studies (Ewas / Envwas) Analysis. SM J Environ Toxicol 7: 4.

The first studies using the EnvWAS approach were conducted in 2010, with two articles introducing this new approach to study the association between the exposome and a health event, namely diabetes [3,4]. This approach aims to study the association between a complex health issue and several environmental exposure factors. In this approach, multiple environmental exposures are analyzed in relation to the same health event, incorporating methods for controlling alpha risk inflation, which are common when multiple tests are performed. The advantage of this approach is that it enables us to study the association of several environmental factors simultaneously in relation to a health event, and thus to identify new statistically significant protective or risk environmental factors for this event. The use of the EnvWAS could lead to a better understanding of the effects of environments with a wider angle of vision (urban, work, etc.) on people's health in relation to certain environmentally sensitive conditions. Moreover, unlike traditional epidemiological studies that focus on one or a few exposures at a time, the EnvWAS adopts a more comprehensive approach by simultaneously considering multiple environmental factors. This allows researchers to explore the cumulative effects and interactions between different exposures, thus providing a more holistic view of the exposome.

The EnvWAS approach and the creation of codes for the analysis and presentation of results can be complex; therefore, a package is useful to simplify the entire process and make it more accessible. The Elja package simplifies the EnvWAS process and yields tabular and graphical results. Recently developed packages on R, not available on CRAN, have also been developed for this purpose (25,26).

Methods

In the Elja package, EnvWAS employs advanced statistical methods, such as linear regression models, logistic regression models, and generalized linear models, to analyze the complex relationships between exposures (independent variables) and health outcomes (dependent variable). The Akaike Information Criterion (AIC) was used to select the most appropriate model by considering both goodness of fit and model complexity [5], the Bonferroni-corrected alpha threshold [6], and the Benjamini-Hochberg False Discovery Rate (FDR) [7] were used to adjust the significance level in hypothesis testing when multiple comparisons are made simultaneously.

Aim of the Package

Simplification of the EnvWAS approach

The first objective of this package is to simplify the implementation of EnvWAS. The analytical approach consists of creating loops that, through an adapted statistical model, analyze the effect of an exposure (independent variable) on a health event (dependent variable). Each exposure is considered for each model using a loop. This package allows the complete EnvWAS process to be carried out by defining at least two elements: the targeted health event in a table (dataframe) containing, among other variables, the exposures whose association is to be tested.

The visual impact of this package

The Elja package offers two types of outputs. A data frame is provided for each variable tested (as well as all the modalities of these variables), including the value of the estimator (odd ratio or coefficients), its 95% confidence interval, the associated p-value, the number of values considered in the model, and the Akaike Information Criterion (AIC) of the model. In addition, two Manhattan plots can be displayed, both with a visual indicator of the alpha threshold of 0.05, the Bonferroni-corrected alpha threshold, and the Benjamini-Hochberg False Discovery Rate (FDR). The first one, representing all the variables in the EnvWAS analysis. The second one, with only significant values.

Implementation and Design

Installation

The Elja package can be installed from GitHub using the devtools [8] package as follows:

```
install.packages("devtools")
library(devtools)
install_github("Displayr/flipPlots")
```

It is also available in the CRAN directory and can be installed as follows.

```
install.packages("Elja")
```

This package is publicly available and compatible with R for Windows, MacOS, and Ubuntu. It also requires the stats [9], base [9], MASS [10] et dplyr [11] packages and for the graphics aspect, these were created using the ggplot2 package [12].

Main features

Using a data frame containing data on the health event and exposure factors to be studied, it is possible to perform an EnvWAS-type analysis by defining the health event of interest that will be used for the analysis within the function specific to the type of model required. The result is represented in the form of a table with the possibility of also having an automatically generated Manhattan plot. Several functions will be added in response to community needs.

The functions included in the package to perform EnvWAS analysis and present the results follow the same approach. The first step is to construct the equations for each model from the dataset and the instructions provided. The second step is to run each model in a loop and store the results. Finally, the results are displayed in a table and, if required, in a graphical form. This package is designed for linear regression, logistic regression, and generalized linear models called ELJAlinear, ELJAlogistic, and ELJAglm, respectively. The functional approach of the function contained in this package is shown in figure 1.

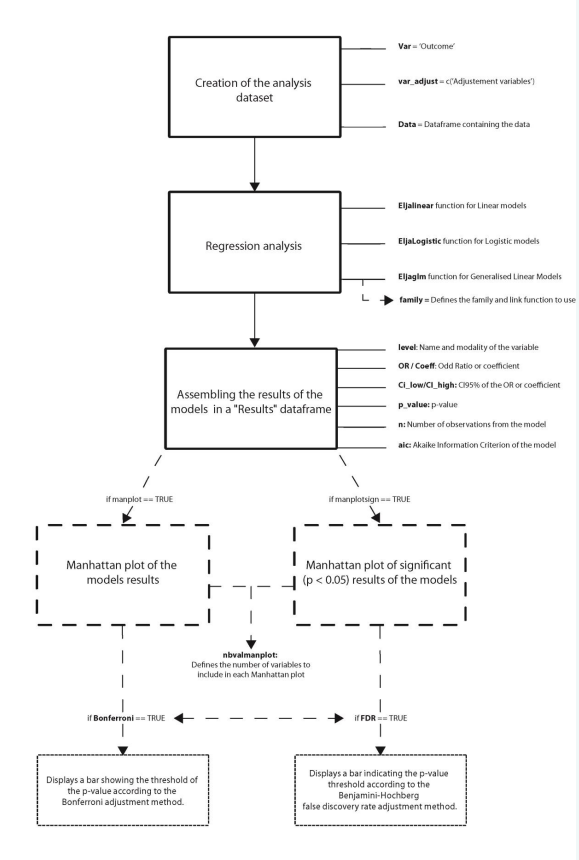

Figure 1 Graphical representation of the functional flow from input to output of the functions in the Elja package.

Application Example

Presentation of the data set

This example used the PIMA dataset [13] extracted from the mlbench package [14]. The PIMA dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The aim of the dataset is to predict whether a patient has diabetes based on certain individual characteristics included in the dataset. This dataset contains only women aged at least 21 years and of Pima Indian origin. This dataset contained the following variables:

- pregnant: Number of times pregnant

- glucose: Plasma glucose concentration a 2 hour in an oral glucose tolerance test
- pressure: Diastolic blood pressure (mm Hg)
- triceps: Triceps skin fold thickness (mm)
- insulin: 2-Hour serum insulin (µU/ml)
- mass: Body mass index (weight in kg/(height in m)^2)
- pedigree: Diabetes pedigree function
- age: Age (years)
- diabetes: test for diabetes (pos/neg)

Use of the Elja Package with PIMA Data

In this example, we will use the "diabetes" variable as a health event and all other variables as exposures. Since the outcome is binary categorical, we use a logistic regression model with the ELJA logistic function. The results of the analysis are presented in table form (table 1) and Manhattan plot (figure 2). A Manhattan plot with significant results is not displayed here.

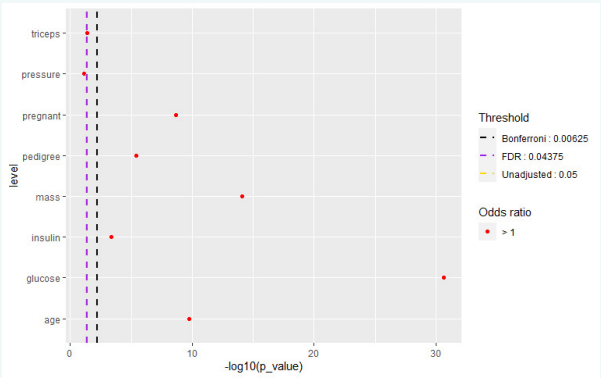

Figure 2 Manhattan plot showing the results of EnvWAS analysis with a logistic regression model on PIMA data, with diabetes as the outcome variable. The p-values are shown on the abscissa. The horizontal bars refer to the different estimated alpha thresholds: in yellow the 5% alpha threshold, in purple the alpha threshold corrected according to Benjamini-Hochberg's False Discovery Rate method, and in black the alpha threshold corrected according to Bonferroni's method.

Table 1: Results of EnvWAS analysis with a logistic regression model on PIMA data with diabetes as the outcome variable. *Akaike information criterion.

Variable	Odd Ratio	95% CI low	95% CI high	*p*-value	n	AIC*
Pregnancies	1.147	1.097	1.200	$2.147.10^{-9}$	768	960.210
Glucose	1.039	1.032	1.045	$2.378.10^{-31}$	768	812.720
Blood Pressure	1.007	0.999	1.016	0.073	768	994.128
Skin Thickness	1.010	1.001	1.019	0.039	768	993.189
Insulin	1.002	1.001	1.004	$4.353.10^{-4}$	768	984.810
BMI	1.098	1.073	1.125	$8.450.10^{-15}$	768	924.714
Diabetes Pedigree Function	2.953	1.880	4.714	$3.703.10^{-6}$	768	974.861
Age	1.043	1.030	1.057	$1.773.10^{-10}$	768	954.720

Conclusion

The Elja package represents a new tool designed to simplify the coding of EnvWAS, and the representation of the results obtained. Its use can enable the analysis of large datasets to determine the impact of environmental factors on health events. We believe that this package can simplify the exploitation of datasets involving several factors related to the same health event. The Elja package is publicly available on CRAN and GitHub.

Funding

This study was funded by the EU-funded URBAN Observatory for Multi-participatory Enhancement of Health and Wellbeing (URBANOME) Project (https://www.urbanome.eu/), EU HORIZON 2020 Grant #945391. Marwan EL HOMSI is a recipient of an EU PhD fellowship within the URBANOME project.

Availability and Requirements

- **Project name:** Elja
- **Project home page:** https://github.com/EHMarwan/Elja
- **Operating system(s):** Windows, MacOS and Ubuntu
- **Programming language:** R
- **Other requirements:** stats, devtools, dplyr, ggplot2, MASS are required.
- **License:** GPL-3

References

1. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. Toxicological Sciences. 2014; 137: 1-2.
2. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. The exposome: molecules to populations. Annu Rev Pharmacol Toxicol. 2019; 59: 107-127.
3. Thomas D. Gene–environment-wide association studies: Emerging approaches. Nat Rev Genet. 2010; 11: 259-272.
4. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLOS ONE. 2010; 5: e10746.
5. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19: 716-723.
6. Dunn OJ. Multiple comparisons among means. Journal of the American Statistical Association. 1961; 56: 52-64.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995; 57: 289-300.
8. Wickham H, Hester J, Chang W, Bryan J. devtools: Tools to Make Developing R Packages Easier. 2023.
9. R Core Team. R: A Language and Environment for Statistical Computing. 2023.
10. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer. 2002.
11. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A Grammar of Data Manipulation. 2023.

12. Wickham H. Ggplot2. Springer International Publishing. 2016.
13. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proc Annu Symp Comput Appl Med Care. 1988: 261-265.
14. Newman D, Hettich S, Blake C, Merz C. UCI Repository of machine learning databases. 1998.