Table of Contents

1. Introduction

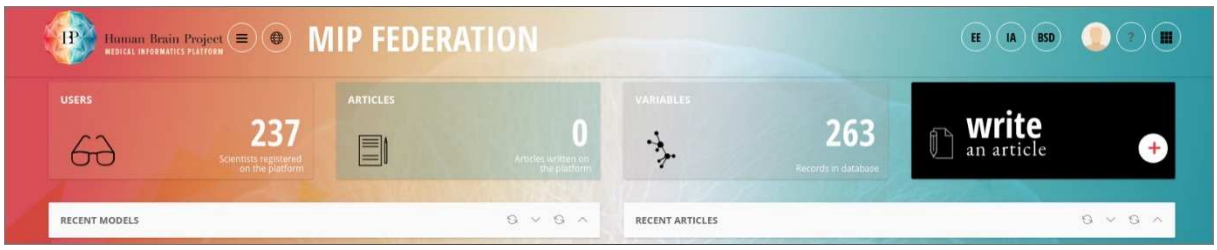
The Medical Informatics Platform (MIP) aims to provide tools to analyse medical data and advance more rapidly in understanding the neurological and psychiatric diseases. The users can access the platform via a Web user interface (UI) (password restricted) where they can run exploratory data analyses, create and share analysis models, execute descriptive statistics, inferential statistics and machine-learning algorithms on user-defined analysis models, as well as collaboratively write articles.

2. MIP User Guidelines

2.1 General Navigation with the MIP

After log-in, the platform opens with the main dashboard.

The dashboard shows a summary of statistics, users, available variables, as well as the latest three saved analysis models and articles (those of the current user or shared among all users). From here, the user may also start writing articles (description of data analyses performed).



At any time the user can return to this page by clicking on the HBP logo on the top left corner.

From the top banner, user can at any time access:

- **My Data**: personal dashboard displaying own work (saved analysis models and articles)
- **My Community**: all work labelled for sharing by any user within its MIP community
- **Functionalities**:
  - Epidemiological Exploration (**EE**)
  - Interactive Analyses (**IA**)
  - Biological Signatures of Diseases (**BSD**)
  - Personal Profile
  - Third-party web applications

2.2 Variables Exploration, Analysis Model and Experiment Design

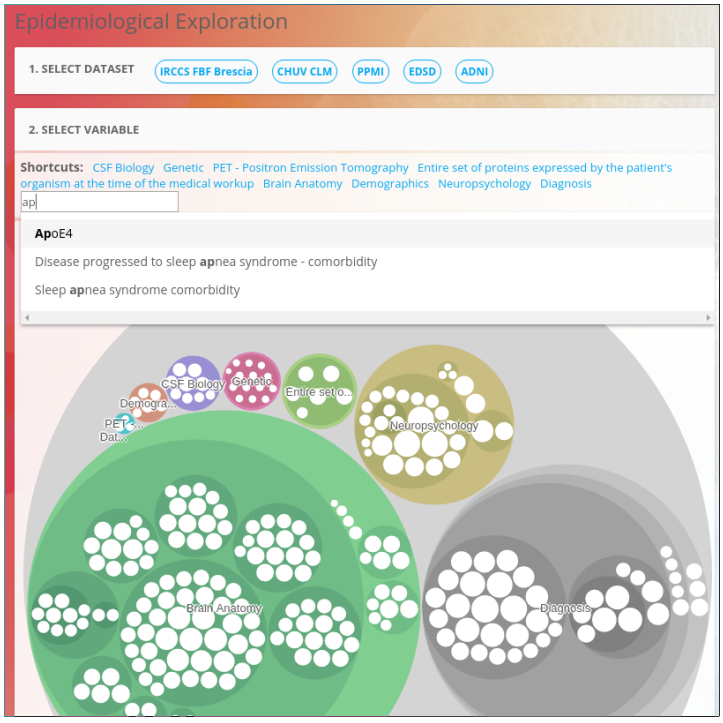
The MIP provides the following functionalities:

- **EE**: Epidemiological Exploration – allows exploration of the available variables, including visualisation of variables' types, selection of variables, visualisation of variables' descriptive information, and definition of analysis models (selection of response and explanatory variables for data analysis);
- **IA**: Interactive Analyses – provides descriptive summary statistics for the defined analysis models, in tabular and graphical formats;
- **BSD**: Biological Signatures of Disease – provides selection and configuration of data analysis algorithms – descriptive and inferential statistics, machine-learning and validation, using analysis models defined with IA functionality.

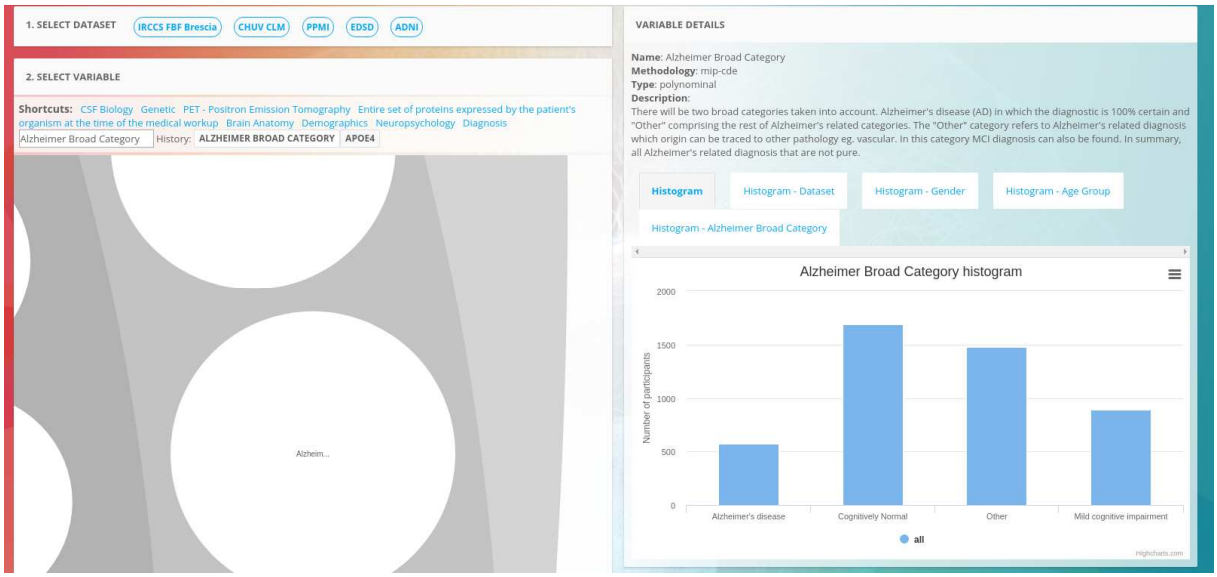
2.2.1 EE: Epidemiological Exploration

The exploration of variables is done through their representation in a circle-pack design. We will call it the MIP Variable Space.

User may also search for a variable by typing its name directly in the search box above the MIP Variable Space (see screen shot below, searching for APOE4 as an example)

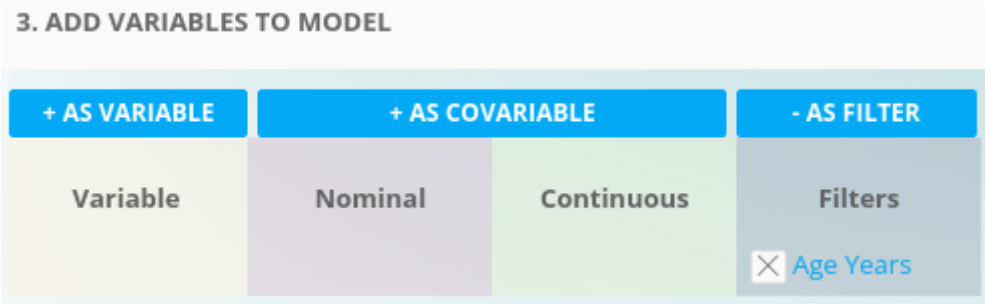


Some variables have descriptive information associated with it. They are displayed in the right-sided panel of the MIP Variable Space (see next screenshot).



User may stop variables exploration here, or continue, to define an analysis model by selecting variables from the MIP Variable Space, as variable (dependent variable), co-variables (independent variables) and filters. It is possible to select a group of variables in one step. See screenshot below.

Variables used for setting up grouping and conditions should be selected as filter variables.



2.2.1.1 Define an Analysis Model

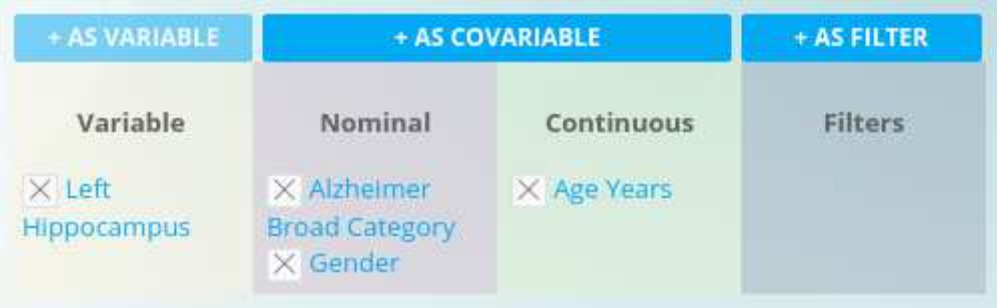
To define an analysis model, user needs to:

1. Search for the variables of interest in the MIP Variable Space, and click on the desired variable, or a group of variables.

Let's take an example: to predict changes in the volume of the Hippocampus in Alzheimer's disease, with respect to age and gender:

- Select the variable Left Hippocampus in the MIP Variable Space and then click on "+ AS VARIABLE" in the model table to define it as a response variable.
- Then select Age Years and Gender variables in EE and click "+ AS COVARIABLES", to define them as explanatory variables. Then do the same for Alzheimer Broad Category. Variables will be automatically classified as "NOMINAL" or "CONTINUOUS" depending on their respective types.

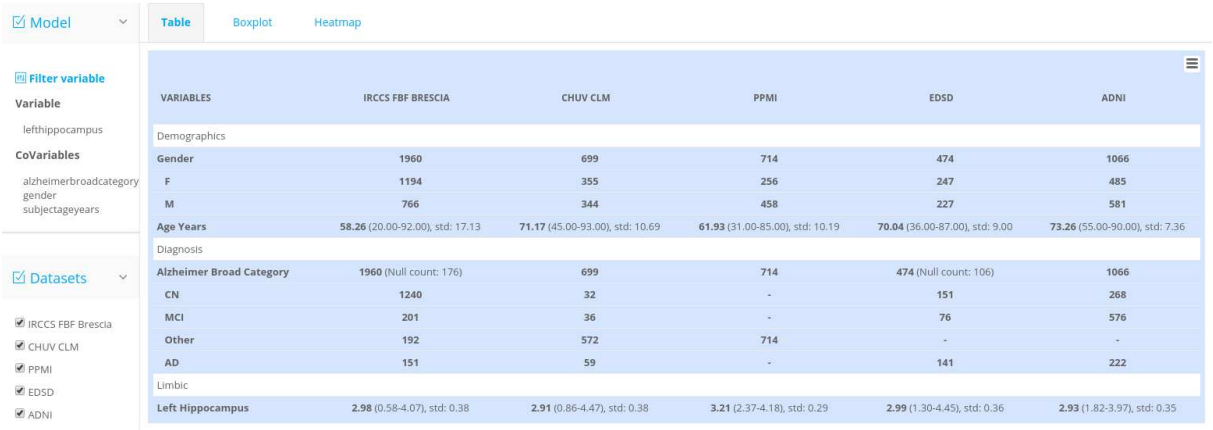
User may at any time remove variables from the Configuration table by clicking the "X" sign. Similarly, user may restore any previously searched variable from the "History" line.



2. When finished with analysis model definition, just click on the Review Model button at the bottom of the screen.

2.2.2 IA: Interactive Analysis

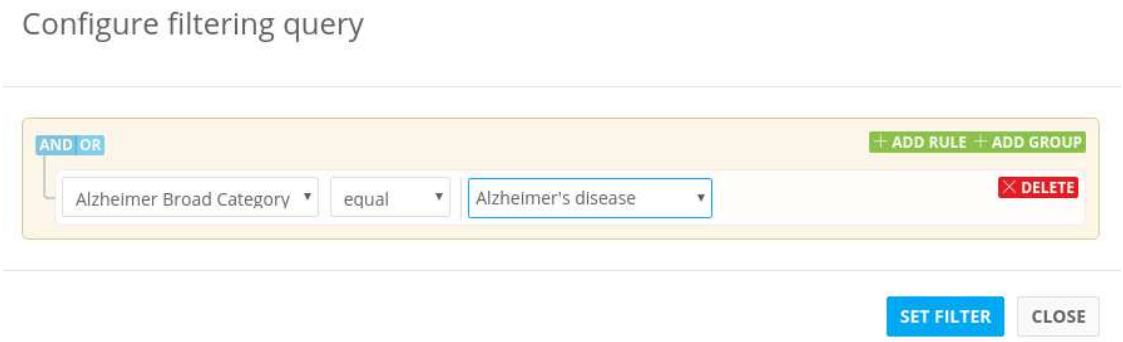
The statistical analysis of the model is run in the MIP. The results are summarized in appropriate tables and visualizations. Statistical description of the defined analysis model is provided in a table and in appropriate visualisations.



VARIABLES	IRCCS FBF BRESCIA	CHUV CLM	PPMI	EDSD	ADNI
Demographics					
Gender	1960	699	714	474	1066
F	1194	355	256	247	485
M	766	344	458	227	581
Age Years	58.26 (20.00-92.00), std: 17.13	71.17 (45.00-93.00), std: 10.69	61.93 (31.00-85.00), std: 10.19	70.04 (36.00-87.00), std: 9.00	73.26 (55.00-90.00), std: 7.36
Diagnosis					
Alzheimer Broad Category	1960 (Null count: 176)	699	714	474 (Null count: 106)	1066
CN	1240	32	-	151	268
MCI	201	36	-	76	576
Other	192	572	714	-	-
AD	151	59	-	141	222
Limbic					
Left Hippocampus	2.98 (0.58-4.07), std: 0.38	2.91 (0.86-4.47), std: 0.38	3.21 (2.37-4.18), std: 0.29	2.99 (1.30-4.45), std: 0.36	2.93 (1.82-3.97), std: 0.35

User can explore the defined analysis model by selecting/deselecting datasets, adding conditions (filters), and selecting visualisations.

To add conditions and grouping, click on the "filter variable" link on the left. That opens a pop-up window for configuring individual rules, their grouping and for selecting their combinations using "and/or" buttons. Conditional and grouping rules are applicable only to variables selected as "filter variables" in the EE functionality. See screenshot below.



After setting up grouping and conditional rules, the analysis model's data is updated. Corresponding tabular and visual summary statistics reflect the analysis model changes.

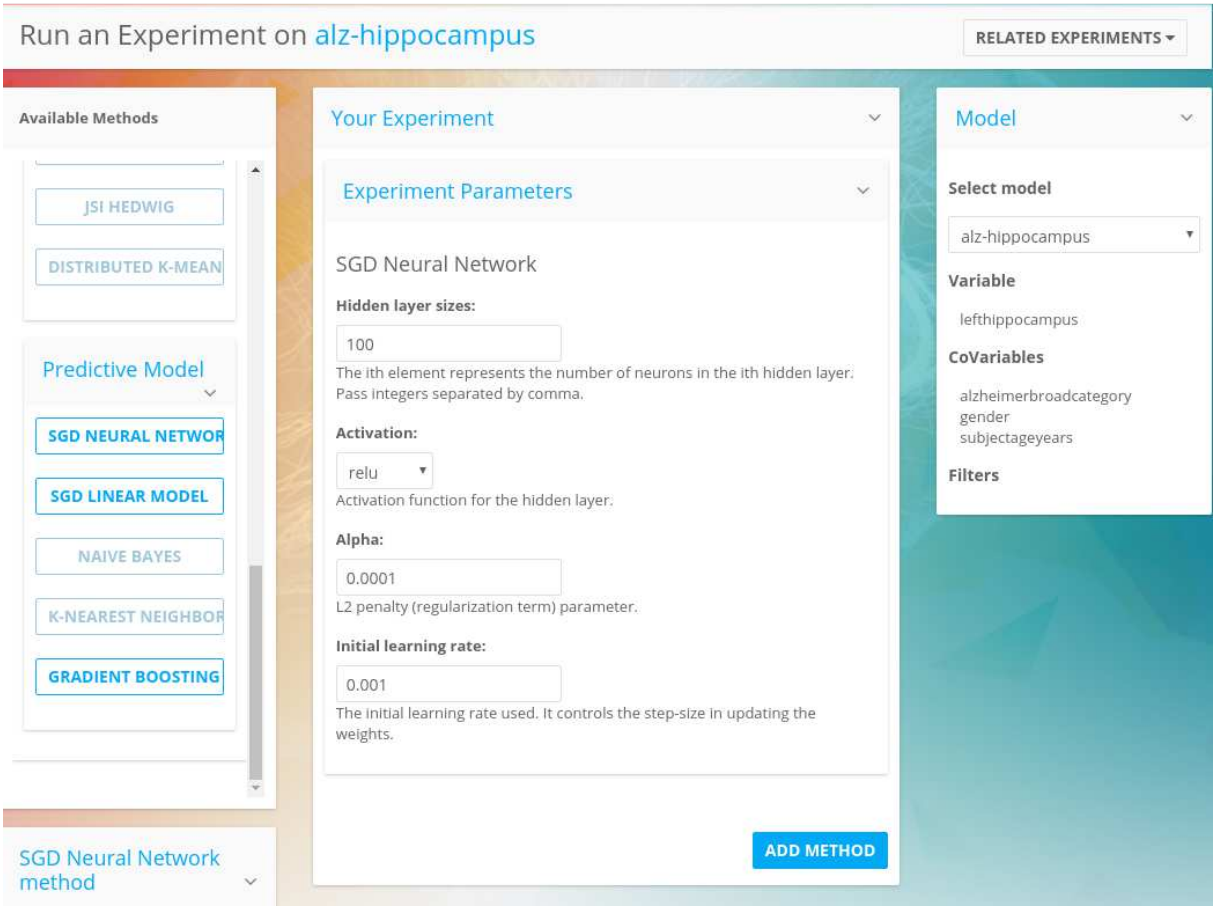
To perform an experiment, user needs to save the analysis model by giving it a name. He can also share it with the MIP community.

2.2.3 BSD: Biological Signatures of Diseases

To configure the experiment:

1- Click "Run Machine Learning Experiment", on the IA screen.

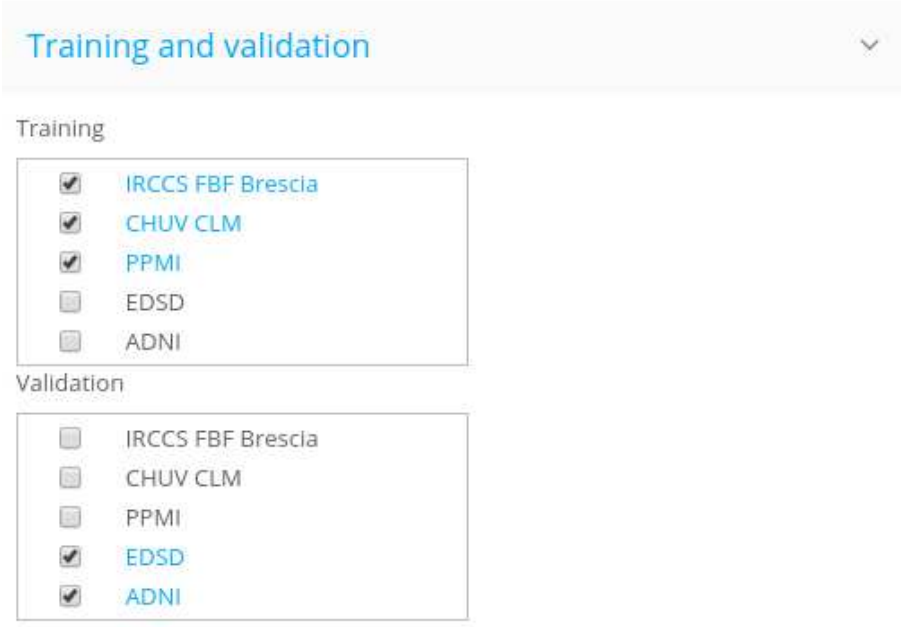
2- In the next BSD screen, choose a method among the "Available Methods", see screenshot below.



3- For some methods user can configure parameters of the algorithm.

4- Click on “Add method” to add it to the methods selected for the experiment. Several methods can be added to an experiment.

5- User can choose to train and validate the experiment on various datasets. See screenshot below.



6- When ready to run the experiment, user needs to give it a name and then to click on RUN EXPERIMENT.

The results of the experiment are presented in textual, tabular or visual format depending on the type of methods chosen and their implementation.

2.2.4 Online Resources

A video demoing the MIP is available on YouTube:
https://www.youtube.com/watch?v=MNWExzouMJw&t=61s

3. Other MIP Functionalities

3.1.1 Writing Articles

1- Go to main MIP Dashboard (click the HBP icon on the top left-hand side);

2- Click on "Write an Article";

3- Write the article: use the editor to add a title, abstract and content to your article;

4- You can also drag and drop results of your models (or others' if are shared) from the left-hand side into the content of your article;

5- Give a name to your article, save it and, optionally, share it.

3.1.2 Accessing my Saved Articles and Models

User can at any time access the already written articles and saved analysis models, via:

- My Data – all own work, shared or unshared;
- My Community – all work shared within the MIP community;

User can preview articles, save them to a file system accessible from his computer or open them for re-editing.

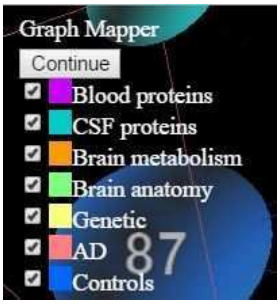
3.1.3 Third Party Applications

3rd Party Application are made by users to provide some insights on specific models or visualisations.

3.1.3.1 3D Biological Rules:

Navigate in a 3D world of variables.

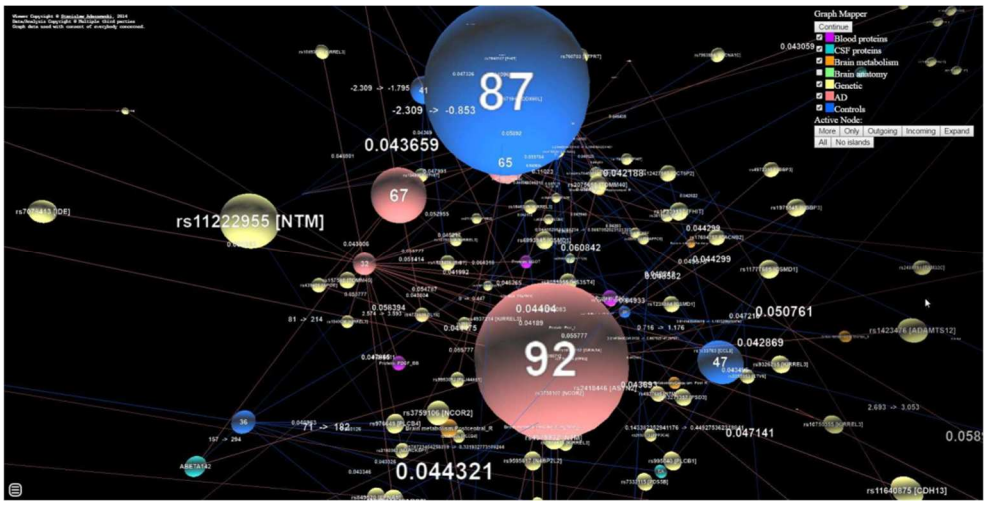
The 3D application shows the results of a rule-based clustering algorithm applied to Alzheimer's disease patients to identify homogeneous subgroups of patients. The hypothesis is that patients in each subgroup have the same underlying cause of the disease. The rule-based algorithm aims to explain the variability between individuals and describes a population by a group of "local over-densities".



These are defined as subspaces over combinations of variables: blood proteins (magenta), CSF proteins (aqua-blue), Brain metabolism (orange), brain anatomy (green)and genetic (yellow).

The red spheres represent AD subgroups and the blue ones healthy controls. The number in a sphere indicates the number of subjects belonging to the subgroup. The edges show rules between the spheres and variables that define each subgroup.

User can also use the left mouse button for rotation, the middle button for zoom and the right mouse button for translation.



User can select and deselect the different variables clicking on the variables tick to make them appear or disappear in the 3D world.

For the genetic variable click “More” to get additional information from the http://www.ensembl.org/ database.

3.1.3.2 2D Biological Rules

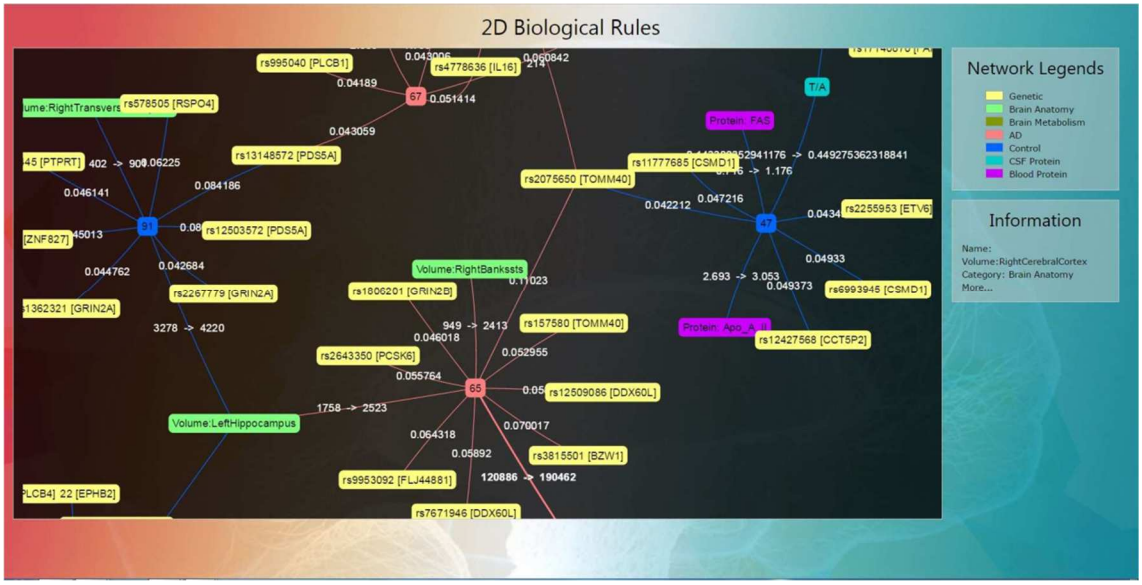
A graphical view of biological rules.

Like the 3D application, 2D application shows the results of a rule-based clustering algorithm applied to Alzheimer’s disease patients to identify homogeneous subgroups of patients. The hypothesis is that patients in each subgroup have the same underlying cause of the disease. The rule-based algorithm aims to explain the variability between individuals and describes a population by a group of “local over-densities”.

These are defined as subspaces over combinations of variables: genetic (yellow), brain anatomy (green), brain metabolism (orange), CSF proteins (aqua-blue) and blood proteins (magenta).

In this application user can select a variable on the 2D Map and get information on the right side.

The red spheres represent AD subgroups and the blue ones healthy controls. The number in a sphere indicates the number of subjects belonging to the subgroup. The edges show the rules between the spheres and the variables that define each subgroup.


2D Biological Rules
Network Legends
Genetic
Brain Anatomy
Brain Metabolism
AD
Control
CSF Protein
Blood Protein
Information
Name:
Volume:RightCerebralCortex
Category: Brain Anatomy
More...
rs995040 [PLCB1]
rs4778636 [IL16]
rs578505 [RSPO4]
rs13148572 [PDS5A]
rs12503572 [PDS5A]
rs2267779 [GRIN2A]
rs1362321 [GRIN2A]
rs1806201 [GRIN2B]
rs2643350 [PCSK6]
rs157580 [TOMM40]
rs2075650 [TOMM40]
rs11777685 [CSMD1]
rs6993945 [CSMD1]
rs2255953 [ETV6]
rs12427568 [CCT5P2]
rs12509086 [DDX60L]
rs3815501 [BZW1]
rs9953092 [FLJ44881]
rs7671946 [DDX60L]
Protein_FAS
Protein_Apo_A_II
Volume:RightBankssts
Volume:LeftHippocampus
T/A
67
91
47
65
0.04189
0.051414
0.043059
0.084186
0.046141
0.042212
0.047216
0.04933
0.044762
0.042684
0.046018
0.052955
0.055764
0.064318
0.070017
0.05892
949 -> 2413
3278 -> 4220
1758 -> 2523
120886 -> 190462
2.693 -> 3.053
0.049373
1.176

4. Glossary

Model: analysis models: Set of variables, co-variables, filters, training and validation datasets.

Epidemiological Exploration (EE): Exploration of variables, their distribution related to datasets, genders, age groups and diagnosis.

Interactive Analyses (IA): Summary statistics of the selected model by dataset.

Biological Signatures of Diseases (BSD): Statistical and machine learning algorithms.

Variable: Dependent or target variable.

Co-variable : Independent variable, or predictor.

- **Nominal**: Categorical type.
- **Continuous**: Numerical (real, integer).

Filter: Include or exclude specifics subjects.

Dataset: Set of patients or subjects related data provided by specific medical institutes.

Federation: Set of algorithms designed to aggregate intermediate results from different datasets.

Distributed: Set of algorithms performed either on different datasets and outputting parallel results, or predictive models trained iteratively on different datasets.

Experiment: Set of algorithms applied on a model.

Methods: Algorithms.

5. Discovering the MIP (Scenarios)

The idea of this chapter is to facilitate learning how to use the MIP, by providing some scripted scenarios.

The scenarios investigate the role of the limbic volumes, like the hippocampus or the entorhinal volumes, in the diagnostic of the Alzheimer disease with the Medical Informatic Platform (MIP).

Scenario 1 = Introduction to the Frontend

After identifying yourself with the identifiers provided by the HBP, the web portal opens on the homepage.

The mouse moves to the upper band, left side

In the upper band of the portal, on the left side, we find two buttons. The left one gives the access to our previous experiments whereas the right one gives access to experiments shared with colleagues.

The mouse moves to the right side

On the right side, there are three main buttons: one to access to the variables exploratory page, a second to access to the model builder page and the third to access to the algorithm page.

Click on the EE button

Let's start with the variables exploratory page. Below the banner, the left part allows us to select data sets and variables, the right part displays the result of our last selection.

The mouse moves over the datasets

In the MIP, there are two types of dataset. Research (academic) datasets like PPMI, ANDI and EDSD, and clinical datasets like CHUV CLM in Lausanne, IRCCS FBF Brescia Italy and CHRU in Lille France. Academic dataset can be used in any local usage, as a reference or to increase the power of the study. By default, all available datasets are selected.

Click on the PPMI button and uncheck the PPMI dataset

Today, as we want to work on dementia, we uncheck the PPMI dataset which is a dataset of Parkinson patients.

The mouse moves down on the variable selection area

Let's now look at the area dedicated to the selection of variables. To facilitate their identification, the variables available in the MIP are grouped into circles. Green circles contain brain volumes, grey circles contain diagnoses, brown circles contain neuropsychological tests, and so on.

Click on the green grey matter circle

The names of the different brain areas appear, and we can easily identify to the limbic area.

Click on the limbic area

Select the Left entorhinal area and as you can see, the right side of the screen is updated.

The mouse moves to the right side

First, we can find some information about the selected variable, its name, acquisition methodology, unit of measurement and type. Below, we have the distribution of all the measures acquired on this variable and stored in the different MIPs, without distinguishing between datasets, gender, age or diagnosis.

Click on the histogram-dataset button and deselect PPMI.

Only only ADNI, EDSD and CLM remain.

By selecting the dataset button, we can observe the distribution of the variable of interest over the different datasets. It shows that ADNI, EDSD and CLM datasets have the same gaussian shape, with the maximum between 1.46 and 1.67 cm3.

Click on the histogram-gender button and deselect PPMI

In the same manner, by selecting the gender button, we can observe that the volume of the left entorhinal area is smaller for women than for men. This is not surprising because generally speaking, women's brains are smaller than men's.

Click on the histogram-Age-group button and deselect -50y

By selecting the gender button, we can observe that the volume of young patient, group 50-59 years have a bigger left entorhinal area than old patients, group +80 years.

Click on the histogram-ABC button and deselect Other and MCI

And finally, the most interesting, all datasets combined, the volume of the left entorhinal area is smaller for Alzheimer patients than for cognitive normal patients. It let us think that this area is a good predictor of the Alzheimer disease.

Scenario 2 = Dataset statistical comparison

Move the mouse down and click as-variable button.

We now go a step forward and check if the volume of the left entorhinal area is statistically different in the research datasets ADNI, EDSD and the clinical dataset CLM. To do that, you need to select both variables, left entorhinal area and dataset and add them to the selection below. Then add left entorhinal area as variable.

The mouse moves to the left in the text area

To select the 'dataset' variable, first search it by writing its name.

The mouse moves to the right side and click on the +as covariable

And add it to the covariable list.

The mouse moves to the review model button

Validate the selection by clicking on the review model button.

We arrive on a new page called Interactive Analysis which allows us to configure variables and datasets according to our needs. On the left side we retrieve our variables of interest, and below the active datasets. We can modify the list of active datasets by clicking on their respective button. In the centre of the screen, in the "table" tab, there is a summary of the

variables by dataset. We have the mean and standard deviation for continuous variables and the size of each level for nominal variables.

Click on the tab "Boxplot"

In the "Boxplot" tab, we have a representation of each experimental condition in the form of a boxplot. For short, this boxplot shows the dispersion of variables, and we observe that in our case, ADNI, EDSD and CLM looks pretty identical.

The mouse moves to the upper right text field.

Enter the model name and click on Save button

To validate this step, we enter a name for this model and save it. Once is done, we could go to the next step, algorithm selection page.

Click on the 'run machine learning experiment'

This new page is named 'Create Experiment'. On the left side of the page, we find a summary of the model containing variables and datasets. On the right, we have a list of algorithms, in blue the algorithms compatible with the selected data and in grey the algorithms present in the MIP but not compatible with the data.

Click on the Anova method

To compare the left entorhinal area volumes in ADNI, EDSD and CLM, we use the ANOVA method. The center of the screen is updated immediately, and ANOVA parameter appears. As we have only one predictor, we don't have to change the value of this parameter.

The mouse moves to the upper right text field.

Enter the experiment name and click on run button.

Now, we save the experiment by giving it a name and run it. It takes few seconds to calculate and display the results. We can see that there is no statistical difference between these three datasets as p-value is 0.1. So, the conclusion of this first short study is that we can use these data sets all together to increase the power of our future studies with the left entorhinal area.

Scenario 3 = Factorial ANOVA

Let's now look at a new problem. We want to see if the difference of the volume of left entorhinal area for Cognitive Normal and Alzheimer Disease patients vary in function of gender. In other word, we want to see if there is an interaction between gender and Alzheimer categories in the ANOVA test. We must create a new model.

Click on the EE button

To do that we have to create a new model by clicking on the EE button on the right of the top banner.

Click on the PPMI button

We deselect the dataset PPMI.

Click on the circle LEA

We select the grey matter volume, then the limbic volumes and the left entorhinal area

Click on the as variable button, right panel

And add it as variable.

Click on the light green and the grey areas

To select the Alzheimer diagnosis, we have to zoom out in the variable display. To do that you \ just have to click outside of the limbic circle, and then click in the grey circle.

Click on the small grey circle, pick up AlzheimerBroadCategory circle and, in the right panel, click on as co-variable

Now, select circle containing diagnosis variable, select Alzheimer broad category and add it as co-variable.

Move to the search text field

Finally, enter the name gender and add it as a second co-variable.

So, we have selected one variable, the left entorhinal area, and two co-variables Alzheimer broad category and Gender.

Click on the Review model

Let's configure and validate the model.

The overview of the selected variables shows that condition 'Other' is only present in the data set CLM. So, we decide to filter this condition.

Click on the filter button

The filter interface allows us to add filters and combine them. We just want to add a filter that exclude the 'Other' condition of the Alzheimer broad category variable.

Click on the first selector button.

When you click on the selector, all the available variables for filter appear. Pick up Alzheimer broad category, and then select 'not equal' and in the last selector, pick up 'Other'.

Click on add rule button and redo the same but with MCI condition

As we want to compare Normal cognitive patients with Alzheimer Disease patients, we must add a new rule to remove 'Mild Cognitive Impairment' condition from the Alzheimer broad category list of condition.

Click on the save button

For the software to take the filter into account, you must save this filter. Once its down, the overview of the variables is updated and the lines corresponding to 'Other' and 'Mild Cognitive Impairment' are filled with 0.

Click on the run experiment button

As before, we select the Anova button, but make sure that the parameter is set to 'factorial', which corresponds to the interaction.

Enter a name in the text field of the experiment and click on run

Name the experiment, save it and run it... The results show that there is a significant effect of 'Alzheimer broad category', which means that the volume of the left entorhinal area is

different between Alzheimer disease patients and Cognitive Normal patients. We show that there is an effect of the 'Gender', M and F are significantly different, which confirms what we observed previously. And finally, there is no Interaction between 'Alzheimer broad category' and 'Gender'. The latter result can be interpreted as meaning that the disease progresses in the same way in men and women.

Scenario 4 = PCA analysis

We know that a large part of the limbic areas is involved in the Alzheimer disease. But there is a lot of variables in this area and they are not all involved in the disease. To find similarities between limbic variables, we use the MIP to do a new analysis using the PCA algorithm.

Click on the EE button

We must create a new model.

Click on the PPMI button

Deselect the dataset PPMI.

Click on the limbic circle and add it as co-variable.

This time, instead of selecting a single variable, select all limbic variables.

Click on the Review model

The main panel displays all the variables and the data sets information.

Click on the 'heatmap' tab

There is a new tab named 'heatmap'. It displays the correlation matrix of all limbic variables.

Move the mouse over the heatmap

We can see for instance that variable 'left hippocampus' is highly correlated with the variable 'right hippocampus'. But overall, this display is not enough to answer our question.

Move the mouse to the model name text field

Save the model to go to the algorithm view. Select the method named PCA, save the experiment and run it. The result is presented in the form of four graphs. The bottom left graph shows us that there are mainly two orthogonal dimensions remaining after the PCA runs. The top left graph shows us that these two dimensions keep 80% of the overall variance of the limbic area. The bottom right graph shows us the distribution of all the limbic measures in this new referential. And finally, the top right graph tells us that Left/right entorhinal areas and left/right hippocampus areas have almost the same coordinates on this new dimensions. This means that they can be gathered for future analysis.

Scenario 5 = Naive Bayes attempt #1

Now, we want to know if the combined knowledge of the volume of the Left entorhinal area and left hippocampus can allow us to identify Alzheimer patient. To do that we use a Naive Bayes algorithm.

Click on the EE button

First create a brand-new model.

Click on the PPMI button

Deselect the dataset PPMI.

Click on the limbic circle

Select all limbic circle to facilitate the access to its variable and add as co-variables the Left entorhinal area and left hippocampus.

Move to the search text field

Then, enter the name Alzheimer broad category in the search text field and add it as a variable.

So, we have selected Alzheimer broad category as variable and two continuous co-variables, the Left entorhinal area and left hippocampus.

Click on the Review model

Click on the 'Review model' button to go to the next page.

Add filter 'Other' and 'MCI'

Use the same filters as when we tried to see the differential effect of gender on Alzheimer's disease. Save them.

Move the mouse to the model name text field

Name the model, save it and go to the algorithm page.

Click on the method

The algorithm we want to use is the Naive Bayes. It is a classifier, that means that it tries to distinguish Alzheimer patient from normal cognitive patient, based only on their volumes of left entorhinal area and hippocampus. Default parameters are classical. Don't modify them.

Move the mouse to the Experiment name text field

Select the method named Naive Bayes, save the experiment and run it. The result displayed in the central panel contain two parts. In the first we can see that the accuracy of the training is around 75% as much as the precision. In the second part we can see the distribution of the prediction. 91 Alzheimer patients are categorized as Cognitive normal, and 126 cognitive normal patient are categorized as Alzheimer…

Scenario 6 = Naive Bayes attempt #2

As the prediction of our model is not good, we decide to add a new predictor, the neuropsychological tests called MOCA. These tests contain some memory assessment and then can help our algorithm to make good decision.

Click on the name of the model

Here, we don't want to create a new model but rather to evolve this one.

On the top panel of the central display, we have the name of the model in blue.

Click on it to go back to the model page

Click explore button

As you click the 'Explore' button you go back to the variables exploratory page.

Mouse over search text field

Here, search the MOCA variable in the appropriate text field, and add it as a co-variable.

Click on the Review model

Then go to the next page by clicking on the review model button.

Click on filters button

Quickly check that the filters remain active and I give a new name to that updated model

Click on run experiment

Now all that remains is to relaunch the Naive Bayes algorithm. The result displayed in the central panel show now an accuracy of the training around 91% as much as the precision. Now if we look the confusion matrix, we can see that only 7 Alzheimer patients are categorized as Cognitive normal, and 17 cognitive normal patient are categorized as Alzheimer.

Scenario 7 = Naive Bayes attempt #3

These results are satisfying, and we want to use our model to make 'real life' predictions. So, we do our first federated analysis based on this model. The main idea of this study is to train the model as we did previously and to use this trained model, on data present in another hospital, such as Brescia for example. In other words, we use data from CLM, ADNI and EDSD to predict the disease in Brescia patients.

Click on the EE button

So, you have now changed to a federated MIP, first create a new model identical to the previous one.

Click on the PPMI button

Deselect the dataset PPMI and CHRU.

Click on the limbic circle

Add the Left entorhinal area and left hippocampus as co-variables, then add the MOCA as co-variable too, and finally in the search text: add Alzheimer broad category as variable and

Click on the Review model

Click on the 'Review model' button to go to the next page.

Add filter 'Other' and 'MCI'

Use the same filters as when we tried to see the differential effect of gender on Alzheimer's disease. Save them.

Move the mouse to the model name text field

Then name the model, save it and go to the algorithm page.

Click on the method

Select the method named Naive Bayes, but instead of saving it right now, modify the values of some parameters. In the bottom of the central panel, we can select data set to use for training and data set to use for remote validation.

Click on Brescia check button in remote-validation

By default, all the data sets are for use in training mode. We check Brescia in the remote-validation area; hence we can evaluate the model on Brescia's data.

Move the mouse to the Experiment name text field

Now save the experiment and run it.

Scenario 8 = K-nn

In this scenario we would like to increase the precision of the model by using a new algorithm with the same experimental protocol (use K-nn algorithm).

Click related 'experiments button'

We don't have to redo the model here, on the right of the top panel we have a button named 'related experiments'. This button give access to all the experiments realized with the current model, and allows user to create a new experiments.

Click Create new experiment

On the left panel we can check the description of the entire model, datasets, variables and filters.

Click K-nearest neighbours

On the right panel we find the method we are looking for.

Select K-nearest neighbours' predictive method.

Move the mouse to Training and kfold

The distribution of datasets between training and remote-validation remains unchanged.

Move the mouse to parameters

Reduce to 2 the number of neighbours as we want to distinguish between 2 levels, Alzheimer disease versus Cognitive normal.

Move the mouse to the Experiment name text field

Then it remains to save the experiment and run it.