Decision Intelligence
for Organizations in
Network Environments

# User Manual

Jacob Stolk PhD

Simone Stolk MPH

July 2021

# Contents

# Introduction

Data analysis with **DIONE** is typically done in four main steps:

1. Load the data to be analysed. In the current version data can be loaded from CSV files and from common statistics data file formats. When loading data from a CSV file, the data file is assumed to have a header record containing variable names and any number of data records containing values of the variables. Variable names and values are separated by commas (,), tabs (\t), semicolons (;), colons(:) or vertical slashes (|) and optionally enclosed in double quotes ("). Each record is delimited by a line break (CRLF).

2. Explore the data using standard statistical tools to get a better intuition of data characteristics, trends, relationships, etc. Each variable can easily be inspected individually, showing plots of its distribution and statistical information. Pairs of variables can easily be compared for a first intuition of possible relationships between variables.

3. Recode data, if needed: for example, transform numerical data to categorical data or group categories of categorical data to new categories.

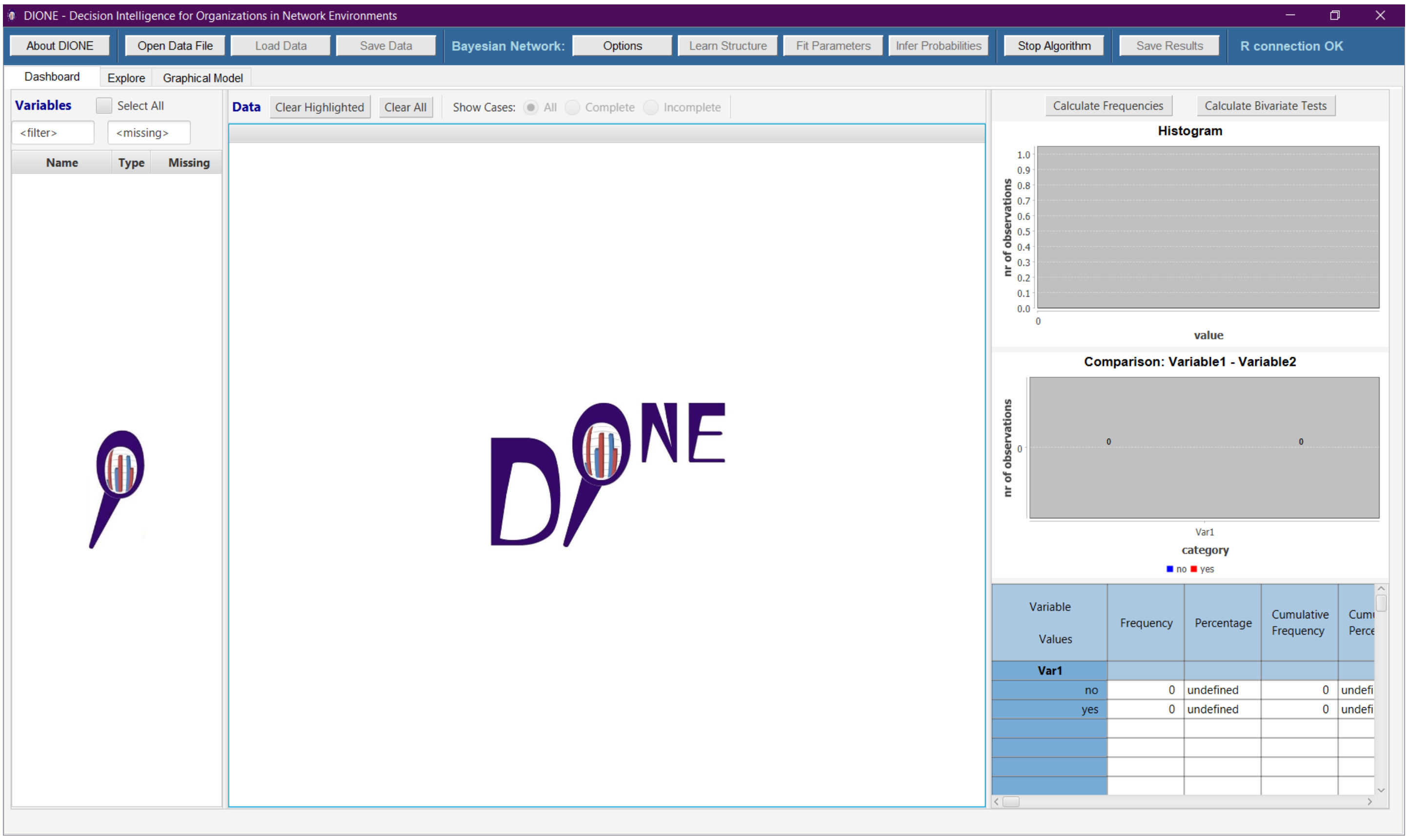
4. Analyse the data using sophisticated data analysis algorithms. At present multivariate analysis is implemented by learning a Bayesian network from the data values of the variables of interest, which will be represented by nodes in the network. The learned network can then be used for inference and for estimating effects of interventions.

5. Save data and/or analysis results. After selecting and/or recoding data for your analysis, you may wish to keep these data to continue your analysis at another time. **DIONE** makes it easy to do this. It is also easy to save analysis results as tables, graphs and charts, ready for your publications.

This guide shows and explains these steps with a simple demonstration data set with fictitious data on variables such as having visited a country with a high incidence of tuberculosis, smoking, age, having the diseases of tuberculosis or lung cancer, and X-ray results. You can of course do similar analyses of your own data.

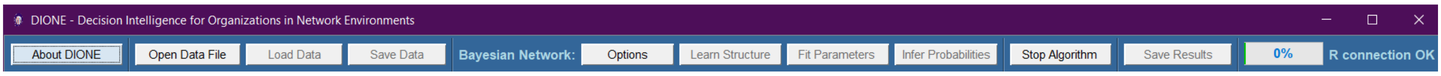
# Start DIONE

To launch **DIONE**, double-click the DIONE desktop icon or select DIONE in the Windows Start Menu . The **DIONE** screen now looks like this:



| Variable Values | Frequency | Percentage | Cumulative Frequency | Cum Perce |
|---|---|---|---|---|
| Var1 | | | | |
| no | 0 | undefined | 0 | undefi |
| yes | 0 | undefined | 0 | undefi |

At the top of the screen there is a toolbar with buttons for loading and saving data and for data analysis. To the right of the buttons there is an area displaying messages about what the program is doing (in light blue), including possible error messages (in red):

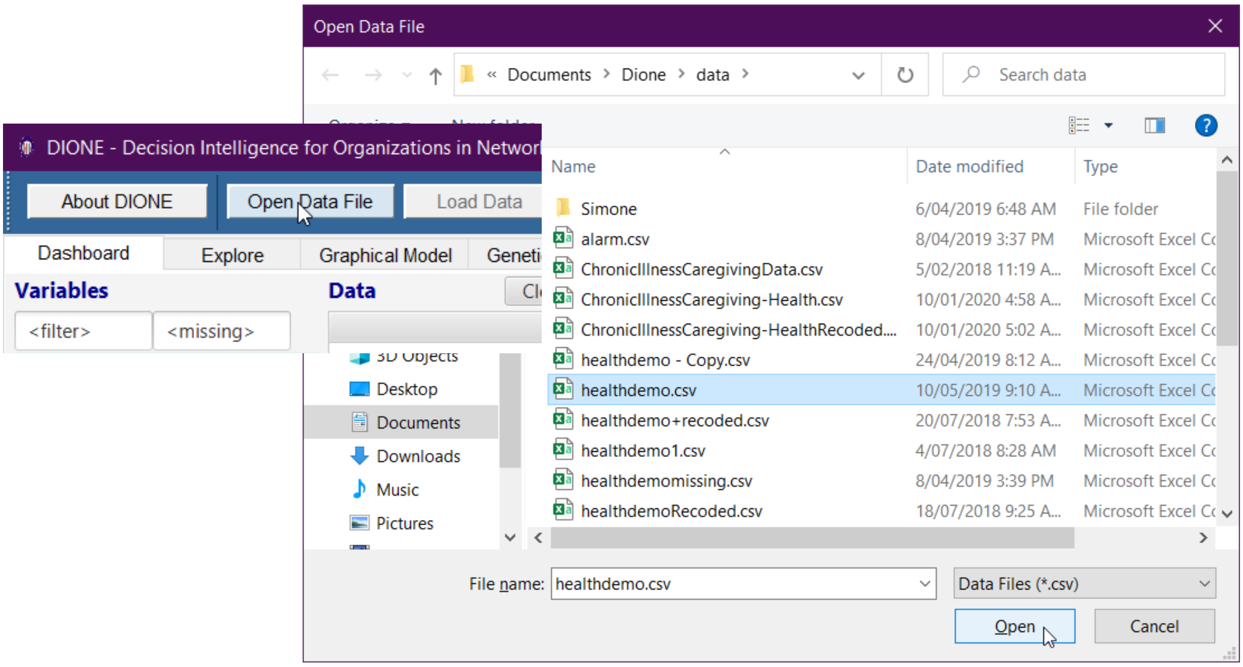


The main area of the screen has three tabs: Dashboard , Explore and Graphical Model . Initially the Dashboard tab is selected.

In the Dashboard tab there are two tables: after loading data the **Variables** table shows the variable names and the **Data** table shows data of selected variables. At the top right-hand side there are visual representations of selected data, and at the bottom right-hand side there is a table with frequency distributions and test results of selected data. In the Explore tab a more detailed view of exploratory data analysis results is shown and there are options to recode data. The Graphical Model tab will show results of Bayesian network learning and inference algorithms.

# Load Data

Data can be loaded from text (`.csv`) files and from common statistics data file formats, including data files with extension `.sav`, `.dta` or `.sas7bdat`. Column headers of a text (`.csv`) data file must be variable names. In this guide the text file `healthdemo.csv` is used. To open a data file, press the Open Data File button to bring up the Open Data File dialog. Now select the data file to analyse and press the Open button:

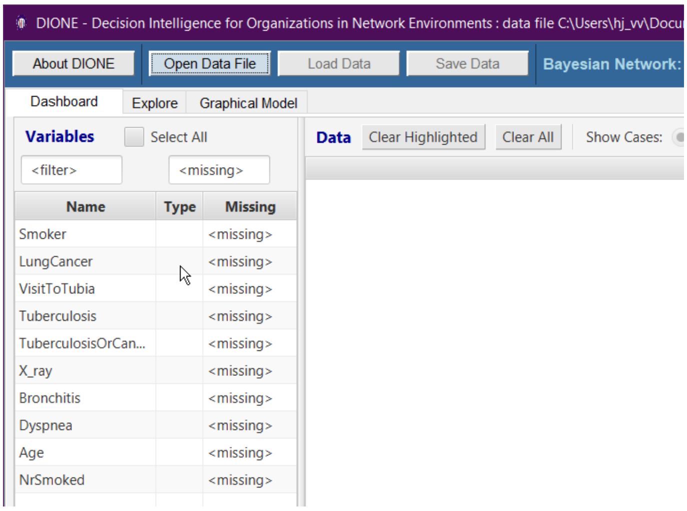


| Name | Date modified | Type |
|---|---|---|
| Simone | 6/04/2019 6:48 AM | File folder |
| alarm.csv | 8/04/2019 3:37 PM | Microsoft Excel Co |
| ChronicIllnessCaregivingData.csv | 5/02/2018 11:19 A... | Microsoft Excel Co |
| ChronicIllnessCaregiving-Health.csv | 10/01/2020 4:58 A... | Microsoft Excel Co |
| ChronicIllnessCaregiving-HealthRecoded.... | 10/01/2020 5:02 A... | Microsoft Excel Co |
| healthdemo - Copy.csv | 24/04/2019 8:12 A... | Microsoft Excel Co |
| healthdemo.csv | 10/05/2019 9:10 A... | Microsoft Excel Co |
| healthdemo+recoded.csv | 20/07/2018 7:53 A... | Microsoft Excel Co |
| healthdemo1.csv | 4/07/2018 8:28 AM | Microsoft Excel Co |
| healthdemomissing.csv | 8/04/2019 3:39 PM | Microsoft Excel Co |
| healthdemoRecoded.csv | 18/07/2018 9:25 A... | Microsoft Excel Co |

File name: healthdemo.csv

Data Files (*.csv)

Now you see the variable names (column headers in the text file) in the **Variables** table at the left of the screen (data values will be loaded in the next step):



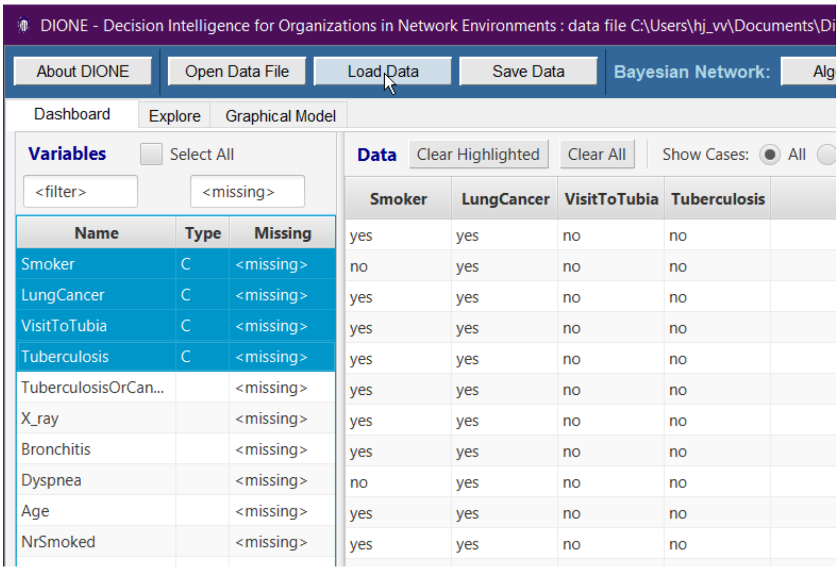
DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Docu

| Name | Type | Missing |
|---|---|---|
| Smoker | | <missing> |
| LungCancer | | <missing> |
| VisitToTubia | | <missing> |
| Tuberculosis | | <missing> |
| TuberculosisOrCan... | | <missing> |
| X_ray | | <missing> |
| Bronchitis | | <missing> |
| Dyspnea | | <missing> |
| Age | | <missing> |
| NrSmoked | | <missing> |

The second column in the **Variables** table is the variables' data type (categorical, ordinal or numerical), to be shown after loading the data values. The third column shows the code for missing values in the data (default is <missing>, but this can be changed if needed).

Select the variables you want to analyse (possibly all) by clicking on their names. You can select variables one at a time, use `Shift-click` or `Ctrl-click` for multiple selection or check ☑ Select All to select all variables.

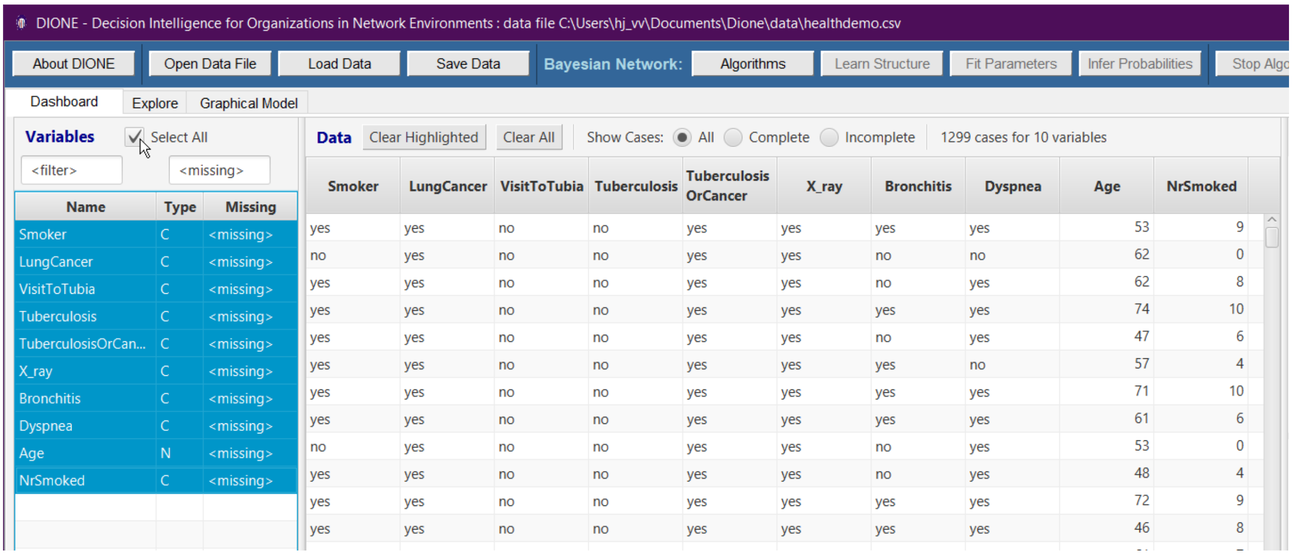
If there are many variables, you can easily locate desired variables by typing part of their names in the <filter> text box (clearing the <filter> text box shows all variables again). When you are satisfied with your selection of variables, press the Load Data button to load the data for the selected variables. You can add more variables and press Load Data again, if desired. You can also press the `Enter` key or double-click to load variables. You can now see data corresponding to the selected variables in the **Data** table in the middle of the screen, for example with 4 variables selected and loaded here:



| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | | <missing> |
| X_ray | | <missing> |
| Bronchitis | | <missing> |
| Dyspnea | | <missing> |
| Age | | <missing> |
| NrSmoked | | <missing> |

| Smoker | LungCancer | VisitToTubia | Tuberculosis |
|---|---|---|---|
| yes | yes | no | no |
| no | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| no | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |

Here all variables have been selected and loaded:

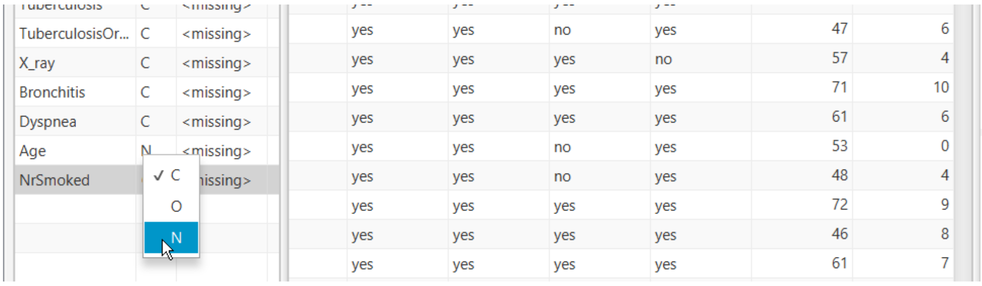


1299 cases for 10 variables

| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | C | <missing> |

| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis | Dyspnea | Age | NrSmoked |
|---|---|---|---|---|---|---|---|---|---|
| yes | yes | no | no | yes | yes | yes | yes | 53 | 9 |
| no | yes | no | no | yes | yes | no | no | 62 | 0 |
| yes | yes | no | no | yes | yes | no | yes | 62 | 8 |
| yes | yes | no | no | yes | yes | yes | yes | 74 | 10 |
| yes | yes | no | no | yes | yes | no | yes | 47 | 6 |
| yes | yes | no | no | yes | yes | yes | no | 57 | 4 |
| yes | yes | no | no | yes | yes | yes | yes | 71 | 10 |
| yes | yes | no | no | yes | yes | yes | yes | 61 | 6 |
| no | yes | no | no | yes | yes | no | yes | 53 | 0 |
| yes | yes | no | no | yes | yes | no | yes | 48 | 4 |
| yes | yes | no | no | yes | yes | yes | yes | 72 | 9 |
| yes | yes | no | no | yes | yes | yes | yes | 46 | 8 |

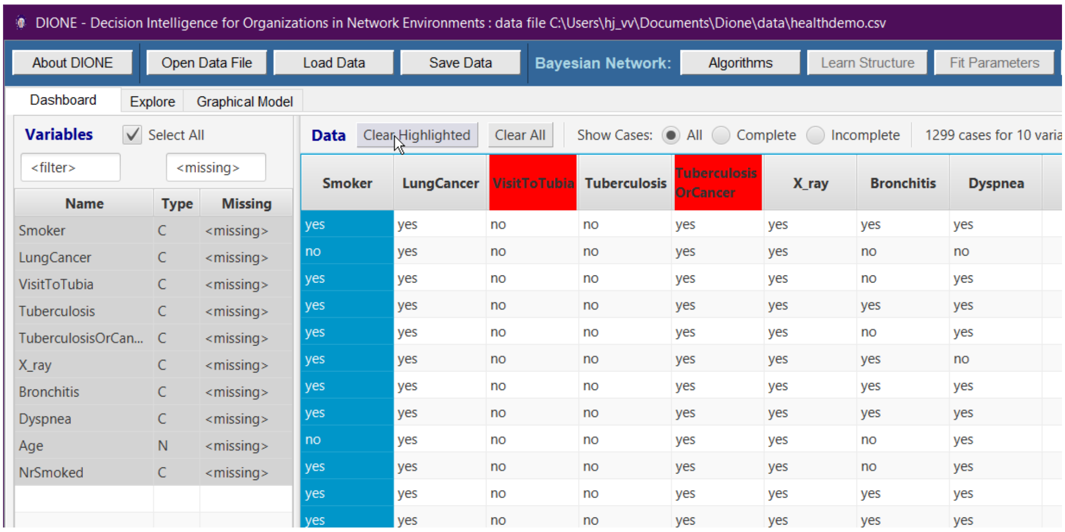
The column **Type** in the **Variables** table shows the data type of the variables, categorical (C), ordinal (O) or numerical (N). **DIONE** infers the data type from the loaded data. If the data for a variable have text values such as `yes` or `no`, the variable is categorical. If the data of a variable consist of numbers, the variable is numerical, unless the number of distinct number values is less than 12, in which case it is assumed the numbers are codes for different levels and the variable is categorical. The data type of a variable can be changed by clicking in the type cell of the variable and selecting a different type, for example making variable **NrSmoked** numerical:



| | | |
|---|---|---|
| TuberculosisOr... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | | <missing> |

| | | | | | | |
|---|---|---|---|---|---|---|
| | yes | yes | no | yes | 47 | 6 |
| | yes | yes | yes | no | 57 | 4 |
| | yes | yes | yes | yes | 71 | 10 |
| | yes | yes | yes | yes | 61 | 6 |
| | yes | yes | no | yes | 53 | 0 |
| | yes | yes | no | yes | 48 | 4 |
| | yes | yes | yes | yes | 72 | 9 |
| | yes | yes | yes | yes | 46 | 8 |
| | yes | yes | yes | yes | 61 | 7 |

If you want to remove variables from the analysis, right-click on the column headers of these variables to highlight them in red. If you change your mind about removing a variable, right-click again on the selected column header. Press button Clear Highlighted to remove the variables highlighted in red:



DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemo.csv

☑ Select All

<filter> <missing>

Show Cases: ● All ○ Complete ○ Incomplete 1299 cases for 10 varia

| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | C | <missing> |

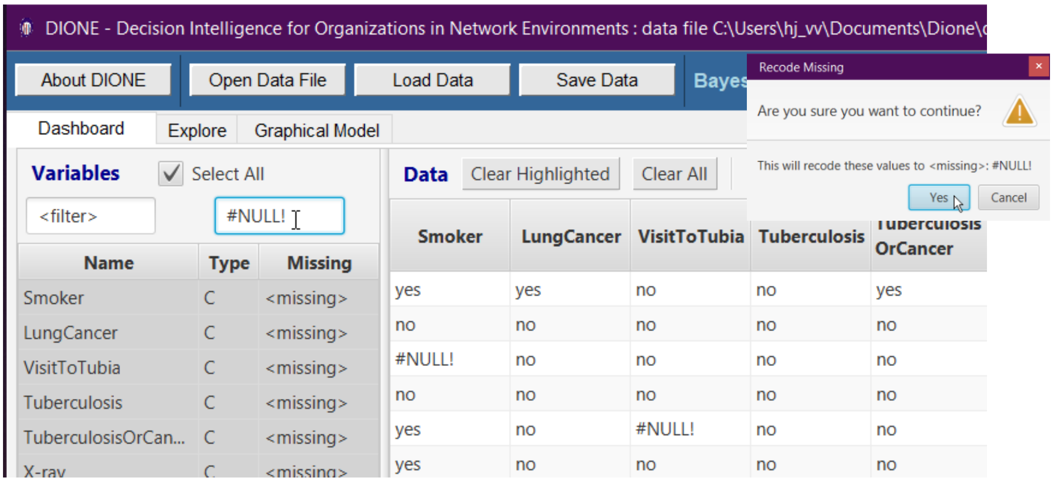
| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis | Dyspnea |
|---|---|---|---|---|---|---|---|
| yes | yes | no | no | yes | yes | yes | yes |
| no | yes | no | no | yes | yes | no | no |
| yes | yes | no | no | yes | yes | no | yes |
| yes | yes | no | no | yes | yes | yes | yes |
| yes | yes | no | no | yes | yes | no | yes |
| yes | yes | no | no | yes | yes | yes | no |
| yes | yes | no | no | yes | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes | yes |
| no | yes | no | no | yes | yes | no | yes |
| yes | yes | no | no | yes | yes | no | yes |
| yes | yes | no | no | yes | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes | yes |

These variables are removed from the **Data** table and will not be used for the Bayesian network analysis but are still loaded. They can be added again to the **Data** table, if desired, by selecting them in the **Variables** table and pressing the `Enter` key. You can also press button Clear All to remove all variables from the **Data** table and start again selecting variables form the **Variables** table.

You can also remove variables from the **Variables** table by selecting them and pressing the `Delete` key to unload them. This way it is easier to remove many variables at a time, but be aware you can <u>no longer access</u> these variables in the current session.

# Missing Values

The default code for missing values is `<missing>`. If your data have a special code for missing values, you can set that code in the <missing> text box by typing the desired code, pressing the `Enter` key and confirming the recoding in the next dialog, as in this example with data from file `healthdemomissing1.csv`:



#NULL!

Recode Missing

Are you sure you want to continue?

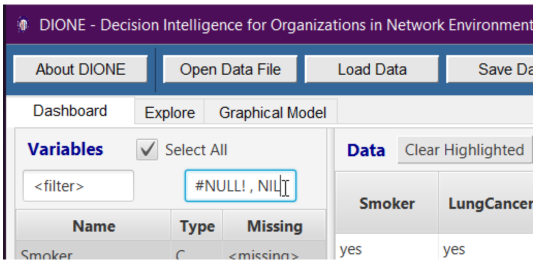
This will recode these values to <missing>: #NULL!

Yes Cancel

| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X-ray | C | <missing> |

| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer |
|---|---|---|---|---|
| yes | yes | no | no | yes |
| no | no | no | no | no |
| #NULL! | no | no | no | no |
| no | no | no | no | no |
| yes | no | #NULL! | no | no |
| yes | no | no | no | no |

For some data more than one value is considered to represent a missing value. You can set several values as missing value by entering them in the <missing> field, separated by commas:

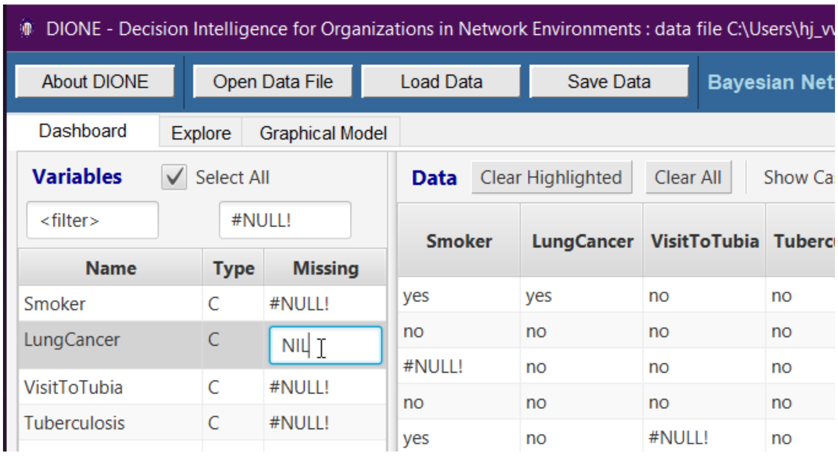


#NULL! , NIL

| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |

| Smoker | LungCancer |
|---|---|
| yes | yes |

You can also set a code for a specific variable in the **Variables** table row for that variable, for example `NIL` for variable **LungCancer** instead of `#NULL!` for all other variables here:



#NULL!

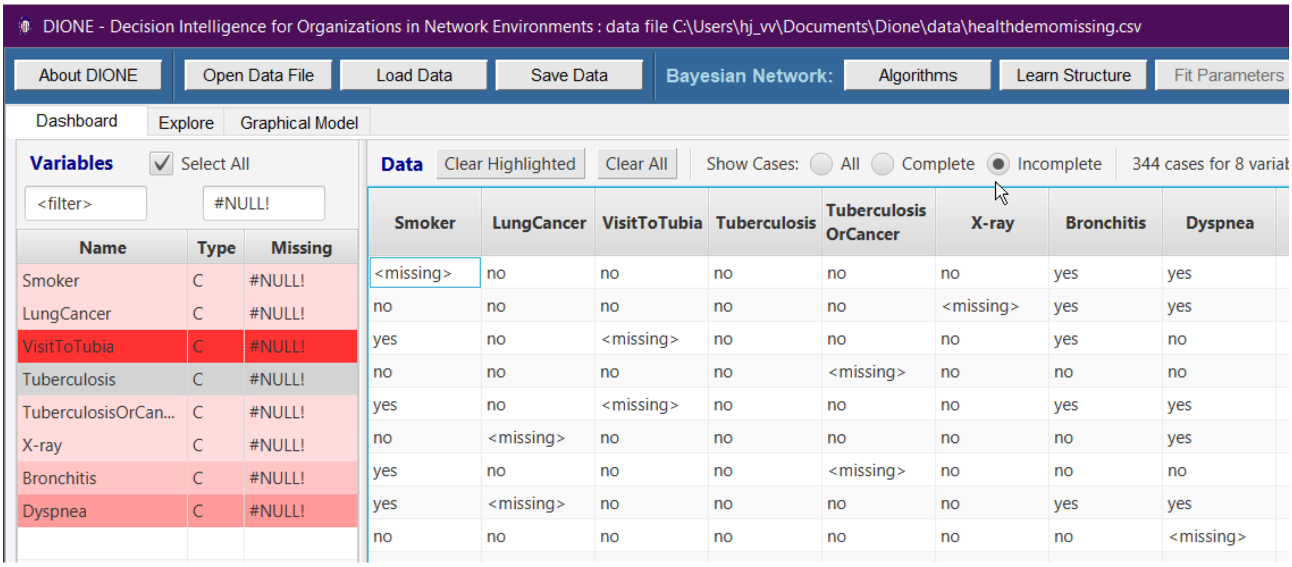
| Name | Type | Missing |
|---|---|---|
| Smoker | C | #NULL! |
| LungCancer | C | NIL |
| VisitToTubia | C | #NULL! |
| Tuberculosis | C | #NULL! |

| Smoker | LungCancer | VisitToTubia | Tuberc |
|---|---|---|---|
| yes | yes | no | no |
| no | no | no | no |
| #NULL! | no | no | no |
| no | no | no | no |
| yes | no | #NULL! | no |

## Complete and Incomplete Cases

Each row of loaded data contains a specific combination of values for the variables that are selected for analysis. We call such a combination of values a case. For many analyses, including common Bayesian network structure learning algorithms, it is necessary to have only complete cases. i.e. cases with no missing values for any of the selected variables. In **DIONE** it is easy to inspect complete and incomplete cases for a set of selected variables.

For example, with the data from file `healthdemomissing.csv` and missing value code set to `#NULL!`, check radio button ○ Incomplete :



DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemomissing.csv

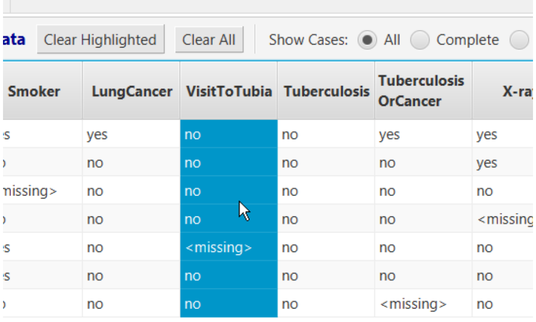
#NULL!

Show Cases: ○ All ○ Complete ● Incomplete 344 cases for 8 variat

| Name | Type | Missing |
|---|---|---|
| Smoker | C | #NULL! |
| LungCancer | C | #NULL! |
| VisitToTubia | C | #NULL! |
| Tuberculosis | C | #NULL! |
| TuberculosisOrCan... | C | #NULL! |
| X-ray | C | #NULL! |
| Bronchitis | C | #NULL! |
| Dyspnea | C | #NULL! |

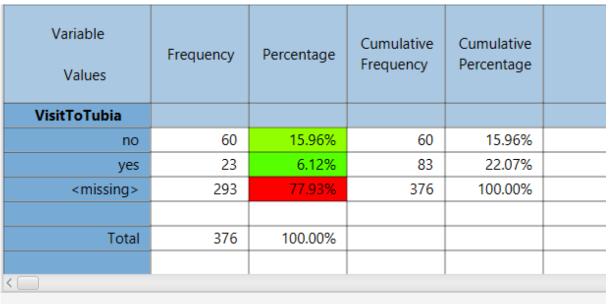
| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X-ray | Bronchitis | Dyspnea |
|---|---|---|---|---|---|---|---|
| <missing> | no | no | no | no | no | yes | yes |
| no | no | no | no | no | <missing> | yes | yes |
| yes | no | <missing> | no | no | no | yes | no |
| no | no | no | no | <missing> | no | no | no |
| yes | no | <missing> | no | no | no | yes | yes |
| no | <missing> | no | no | no | no | no | yes |
| yes | no | no | no | <missing> | no | no | no |
| yes | <missing> | no | no | no | no | yes | yes |
| no | no | no | no | no | no | no | <missing> |

The **Data** table now shows the cases with missing values, with missing value codes replaced by the standard code `<missing>` used by **DIONE**. To give more insight in numbers of missing values of all variables, the variables in the **Variables** table are colour-coded in shades of red: darker red means a higher percentage of missing values, so in this example variable **VisitToTubia** has the highest proportion of missing values. Values of a variable can be inspected by double-clicking in the variable's data column in the **Data** table, with radio button ● All selected to see all values:



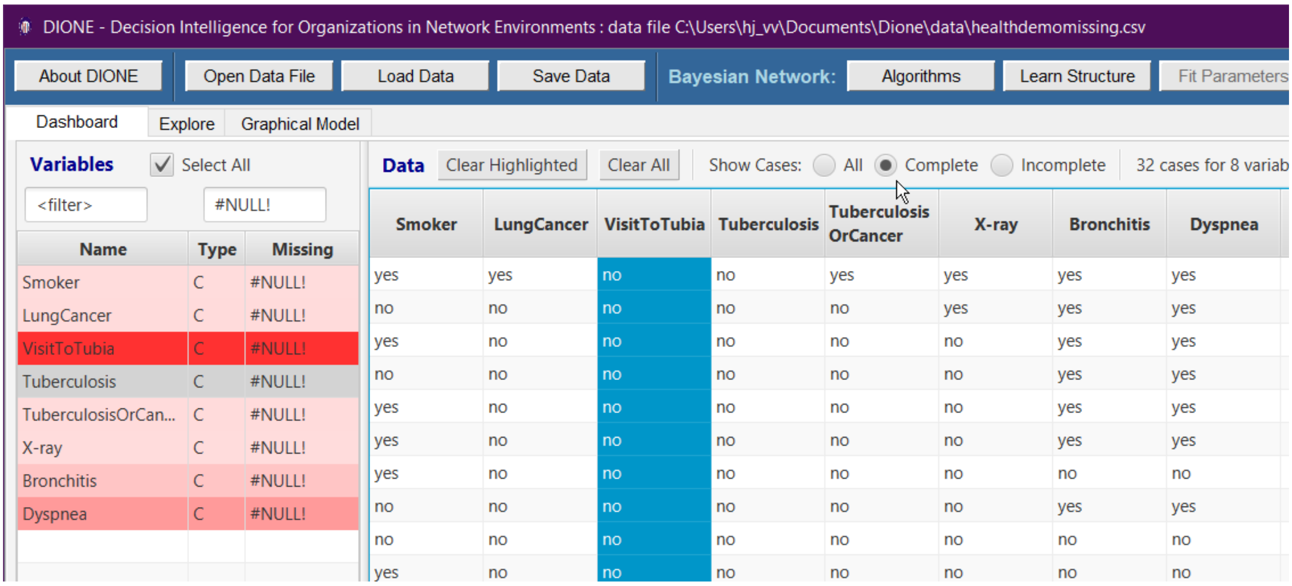
Show Cases: ● All ○ Complete ○

| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X-ra |
|---|---|---|---|---|---|
| s | yes | no | no | yes | yes |
| ) | no | no | no | no | yes |
| nissing> | no | no | no | no | no |
| ) | no | no | no | no | <missing |
| s | no | <missing> | no | no | no |
| s | no | no | no | no | no |
| ) | no | no | no | <missing> | no |



| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| **VisitToTubia** | | | | |
| no | 60 | 15.96% | 60 | 15.96% |
| yes | 23 | 6.12% | 83 | 22.07% |
| <missing> | 293 | 77.93% | 376 | 100.00% |
| | | | | |
| Total | 376 | 100.00% | | |

To show complete cases, select radio button ○ Complete :



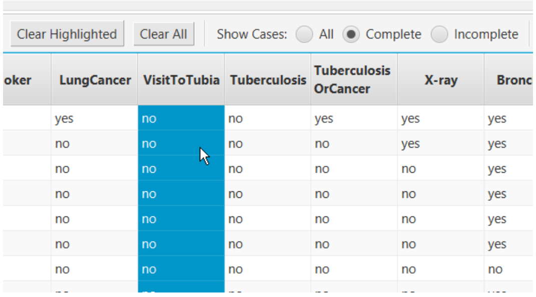
DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemomissing.csv

Show Cases: ○ All ● Complete ○ Incomplete   32 cases for 8 variab

| Name | Type | Missing |
|---|---|---|
| Smoker | C | #NULL! |
| LungCancer | C | #NULL! |
| VisitToTubia | C | #NULL! |
| Tuberculosis | C | #NULL! |
| TuberculosisOrCan... | C | #NULL! |
| X-ray | C | #NULL! |
| Bronchitis | C | #NULL! |
| Dyspnea | C | #NULL! |

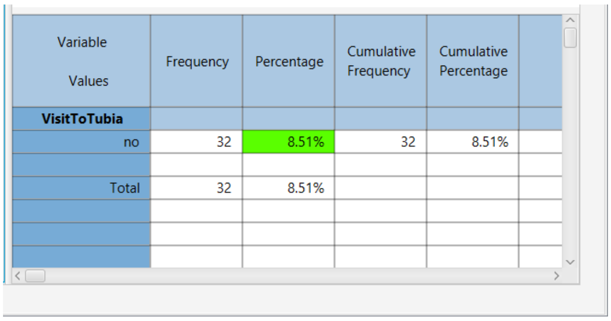
| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X-ray | Bronchitis | Dyspnea |
|---|---|---|---|---|---|---|---|
| yes | yes | no | no | yes | yes | yes | yes |
| no | no | no | no | no | yes | yes | yes |
| yes | no | no | no | no | no | yes | yes |
| no | no | no | no | no | no | yes | yes |
| yes | no | no | no | no | no | yes | yes |
| yes | no | no | no | no | no | yes | yes |
| yes | no | no | no | no | no | no | no |
| no | no | no | no | no | no | yes | yes |
| no | no | no | no | no | no | no | no |
| yes | no | no | no | no | no | no | no |

You can now also see details on variables for complete cases only by double-clicking in the variable's data column in the **Data** table with the ○ Complete radio button selected:



Show Cases: ○ All ● Complete ○ Incomplete

| oker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X-ray | Bronc |
|---|---|---|---|---|---|---|
| | yes | no | no | yes | yes | yes |
| | no | no | no | no | yes | yes |
| | no | no | no | no | no | yes |
| | no | no | no | no | no | yes |
| | no | no | no | no | no | yes |
| | no | no | no | no | no | yes |
| | no | no | no | no | no | no |



| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| **VisitToTubia** | | | | |
| no | 32 | 8.51% | 32 | 8.51% |
| | | | | |
| Total | 32 | 8.51% | | |

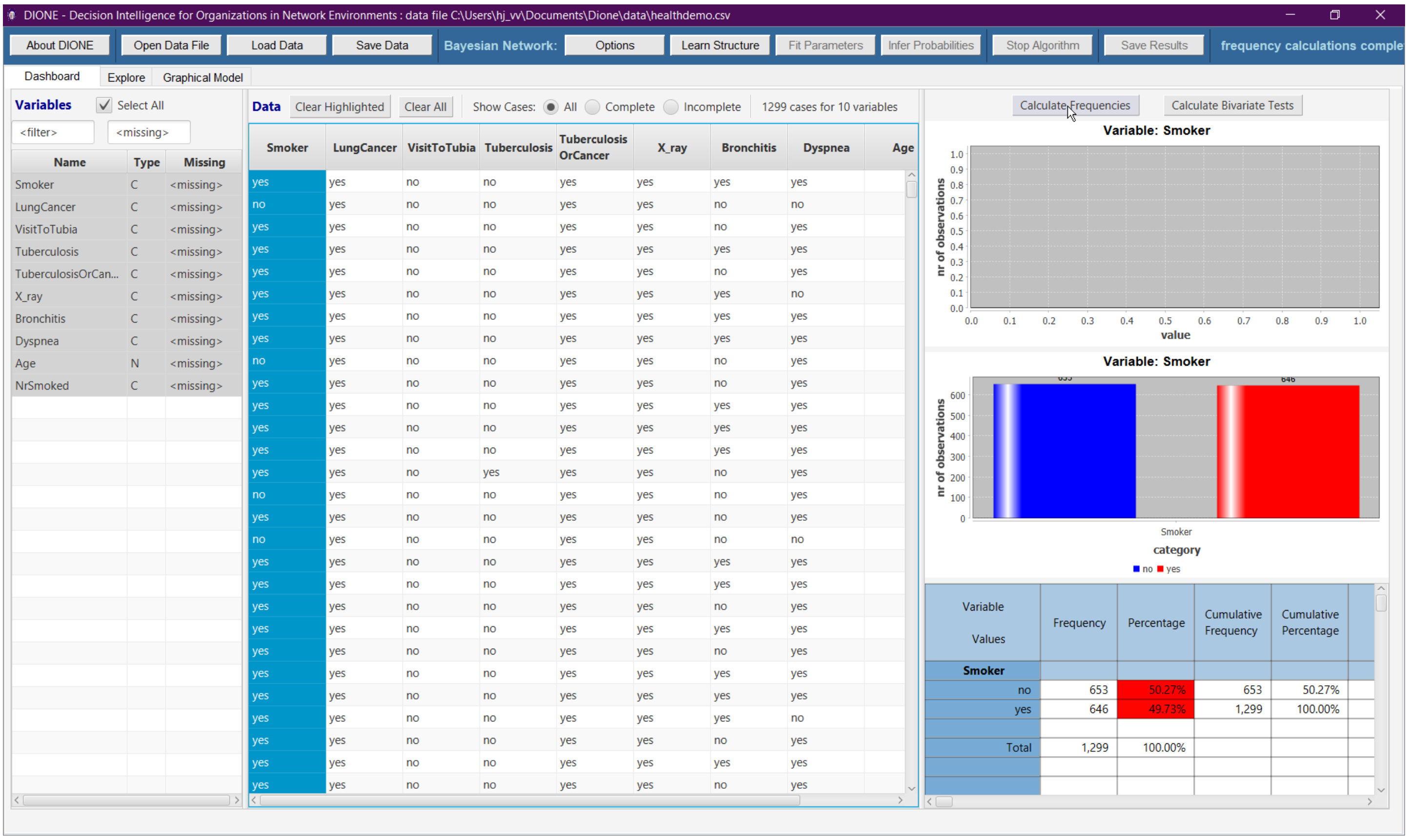
Please note that, for numerical variables that have some nonnumerical values, the nonnumerical values will de recoded to missing when selecting radio button ○ Complete or ○ Incomplete .

# Explore Data

Before proceeding to sophisticated data analysis, it is usually a good idea to explore the data with relatively simple tools to develop a better understanding and intuition of their characteristics. With **DIONE** it is simple to explore data by looking at one variable at a time or by comparing two variables (univariate and bivariate methods): frequency tables, bar charts, histograms, scatter plots and calculation of statistics such as mean, median, mode, standard deviation, chi-square tests and risk ratios with p-values, as well as Bayes factors and correlations. To make it easy to decide which variables are most interesting for multivariate analysis, it is also possible to obtain bivariate comparisons between a variable of interest and all other variables and see in one table which variables are likely to be related to the variable of interest. The variable of interest will in many cases be an outcome variable for which you would like to know causes.

## One Variable

Select a variable in the **Data** table, by clicking anywhere in its data column, and press button `Calculate Frequencies`, press the `Enter` key or double-click to see a bar chart and/or histogram and a frequency table of this variable's data at the right of the screen, for example variable **Smoker** with possible values `yes` and `no`:



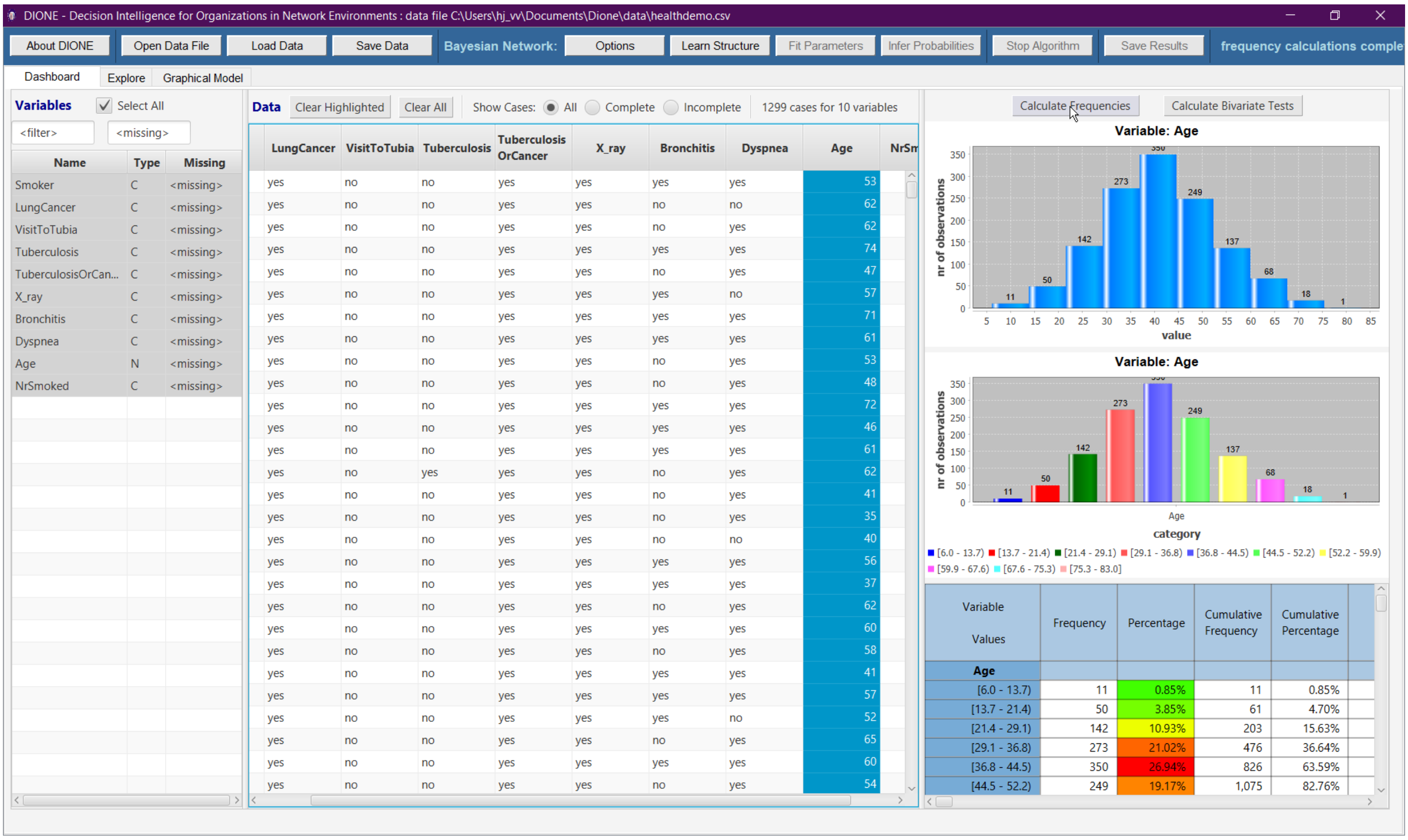
| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | C | <missing> |

| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis | Dyspnea | Age |
|---|---|---|---|---|---|---|---|---|
| yes | yes | no | no | yes | yes | yes | yes | |
| no | yes | no | no | yes | yes | no | no | |
| yes | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | yes | no | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| no | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | yes | yes | yes | no | yes | |
| no | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | no | yes | |
| no | yes | no | no | yes | yes | no | no | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | yes | no | |
| yes | yes | no | no | yes | yes | no | yes | |
| yes | yes | no | no | yes | yes | yes | yes | |
| yes | yes | no | no | yes | yes | no | yes | |

| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| **Smoker** | | | | |
| no | 653 | 50.27% | 653 | 50.27% |
| yes | 646 | 49.73% | 1,299 | 100.00% |
| | | | | |
| Total | 1,299 | 100.00% | | |

For categorical or ordinal variables, a bar chart is shown. For numerical variables a histogram is shown and under the frequency data in the frequency table their mean, median and standard deviation are shown.

This example from the same data set also shows a histogram for a numerical variable: **Age** (a bar chart is also shown to visualize the raw data):



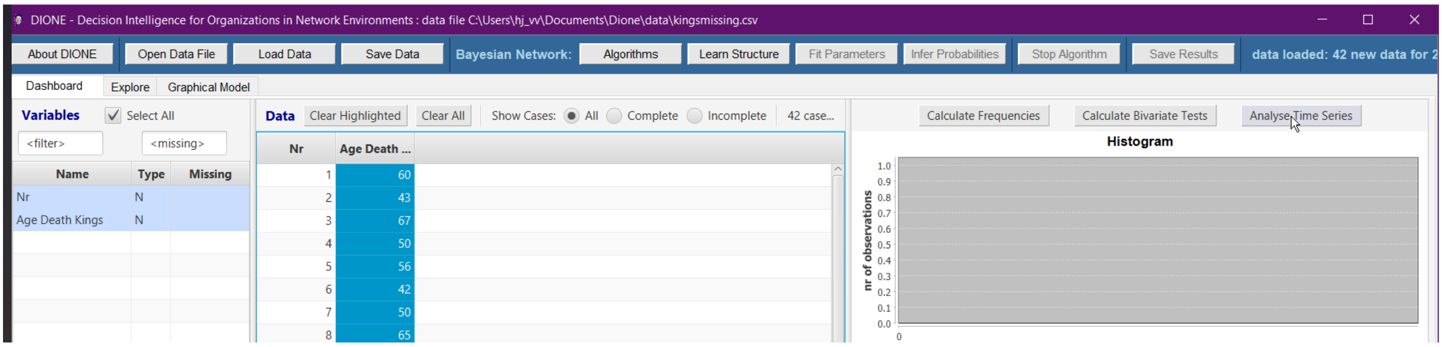
| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | C | <missing> |

| LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis | Dyspnea | Age | NrSm |
|---|---|---|---|---|---|---|---|---|
| yes | no | no | yes | yes | yes | yes | 53 | |
| yes | no | no | yes | yes | no | no | 62 | |
| yes | no | no | yes | yes | no | yes | 62 | |
| yes | no | no | yes | yes | yes | yes | 74 | |
| yes | no | no | yes | yes | no | yes | 47 | |
| yes | no | no | yes | yes | yes | no | 57 | |
| yes | no | no | yes | yes | yes | yes | 71 | |
| yes | no | no | yes | yes | yes | yes | 61 | |
| yes | no | no | yes | yes | no | yes | 53 | |
| yes | no | no | yes | yes | no | yes | 48 | |
| yes | no | no | yes | yes | yes | yes | 72 | |
| yes | no | no | yes | yes | yes | yes | 46 | |
| yes | no | no | yes | yes | yes | yes | 61 | |
| yes | no | yes | yes | yes | no | yes | 62 | |
| yes | no | no | yes | yes | no | yes | 41 | |
| yes | no | no | yes | yes | no | yes | 35 | |
| yes | no | no | yes | yes | no | no | 40 | |
| yes | no | no | yes | yes | yes | yes | 56 | |
| yes | no | no | yes | yes | yes | yes | 37 | |
| yes | no | no | yes | yes | no | yes | 62 | |
| yes | no | no | yes | yes | yes | yes | 60 | |
| yes | no | no | yes | yes | no | yes | 58 | |
| yes | no | no | yes | yes | yes | yes | 41 | |
| yes | no | no | yes | yes | yes | yes | 57 | |
| yes | no | no | yes | yes | yes | no | 52 | |
| yes | no | no | yes | yes | no | yes | 65 | |
| yes | no | no | yes | yes | yes | yes | 60 | |
| yes | no | no | yes | yes | no | yes | 54 | |

| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| **Age** | | | | |
| [6.0 - 13.7) | 11 | 0.85% | 11 | 0.85% |
| [13.7 - 21.4) | 50 | 3.85% | 61 | 4.70% |
| [21.4 - 29.1) | 142 | 10.93% | 203 | 15.63% |
| [29.1 - 36.8) | 273 | 21.02% | 476 | 36.64% |
| [36.8 - 44.5) | 350 | 26.94% | 826 | 63.59% |
| [44.5 - 52.2) | 249 | 19.17% | 1,075 | 82.76% |

## Time Series Data

It is also possible to analyse time series data, including smoothing of a time series by calculating a moving average and decomposition of a time series in trend, periodic and random components. As an example we analyse these data on the age of death of successive kings of England by selecting variable **Age Death Kings** in the **Data** table and clicking button Analyse Time Series :



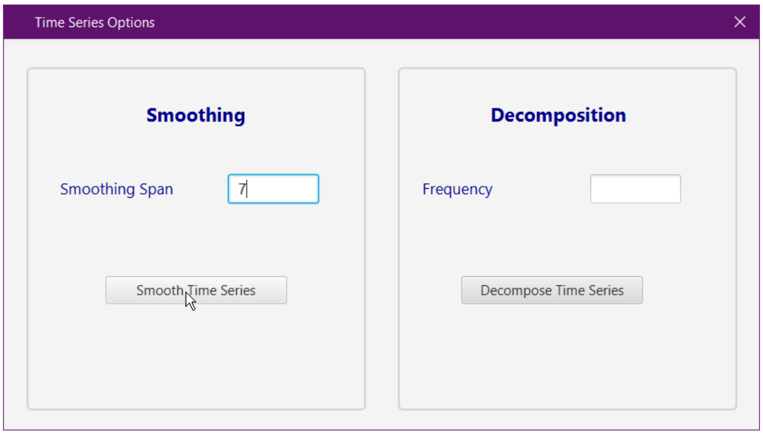
DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\kingsmissing.csv

data loaded: 42 new data for 2

| Name | Type | Missing |
|---|---|---|
| Nr | N | |
| Age Death Kings | N | |

| Nr | Age Death ... |
|---|---|
| 1 | 60 |
| 2 | 43 |
| 3 | 67 |
| 4 | 50 |
| 5 | 56 |
| 6 | 42 |
| 7 | 50 |
| 8 | 65 |

This shows the following dialog. To smooth the time series, set the Smoothing Span parameter, which is the number of values to use for a moving average calculation, form example 7 here:

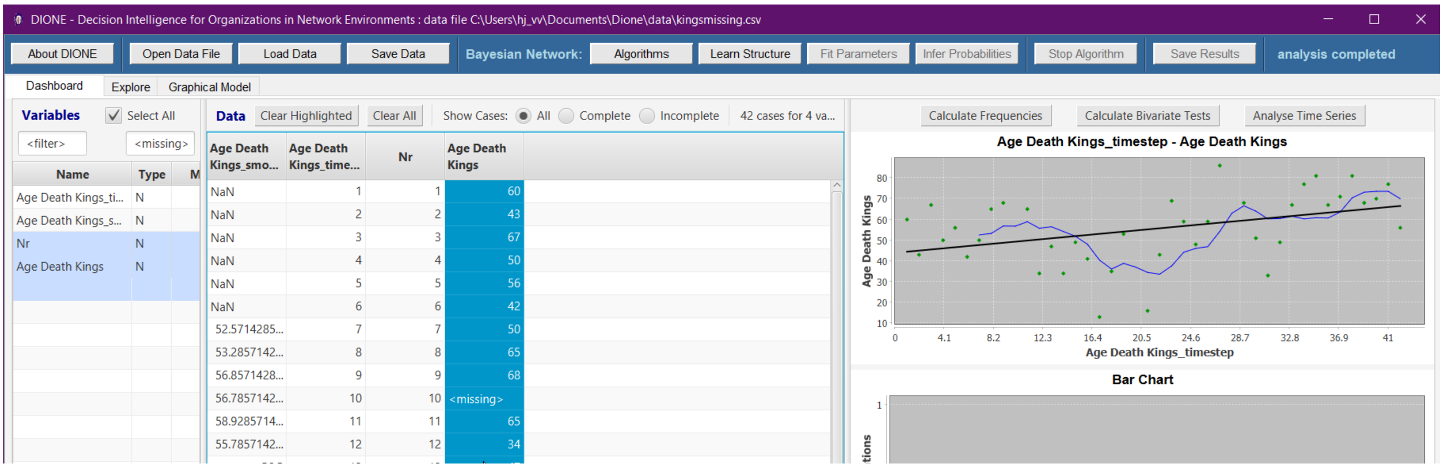


Time Series Options

Smoothing Span: 7

Frequency:

The result is shown here, with data points as dots, the moving average as a blue line and a regression line as a black line:



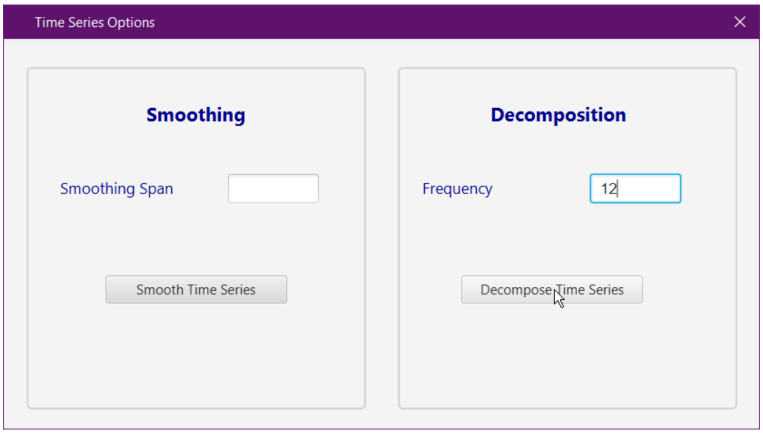
DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\kingsmissing.csv

analysis completed

| Name | Type | M |
|---|---|---|
| Age Death Kings_ti... | N | |
| Age Death Kings_s... | N | |
| Nr | N | |
| Age Death Kings | N | |

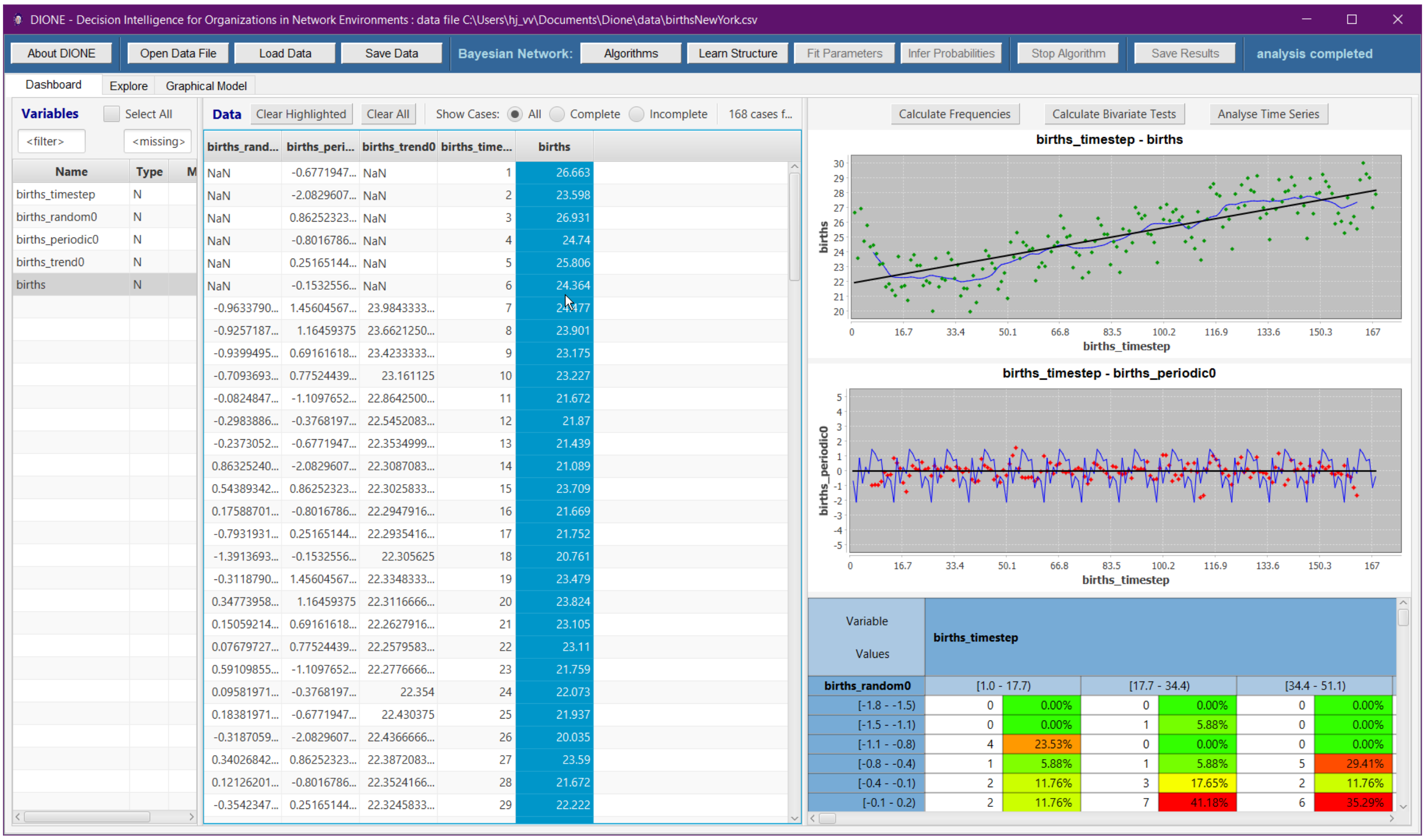
| Age Death Kings_smo... | Age Death Kings_time... | Nr | Age Death Kings |
|---|---|---|---|
| NaN | 1 | 1 | 60 |
| NaN | 2 | 2 | 43 |
| NaN | 3 | 3 | 67 |
| NaN | 4 | 4 | 50 |
| NaN | 5 | 5 | 56 |
| NaN | 6 | 6 | 42 |
| 52.5714285... | 7 | 7 | 50 |
| 53.2857142... | 8 | 8 | 65 |
| 56.8571428... | 9 | 9 | 68 |
| 56.7857142... | 10 | 10 | <missing> |
| 58.9285714... | 11 | 11 | 65 |
| 55.7857142... | 12 | 12 | 34 |

To decompose a time series, set the `Frequency` parameter, which is the number of values assumed for one period of the time series, for example 12 here for a dataset of monthly births in New York:



Frequency: 12

The result is shown here, with the top chart showing again data points as dots, the trend component as a blue line and a regression line as a black line, and the bottom chart showing the random component as dots and the periodic component as a blue line:



| Name | Type | M |
|---|---|---|
| births_timestep | N | |
| births_random0 | N | |
| births_periodic0 | N | |
| births_trend0 | N | |
| births | N | |

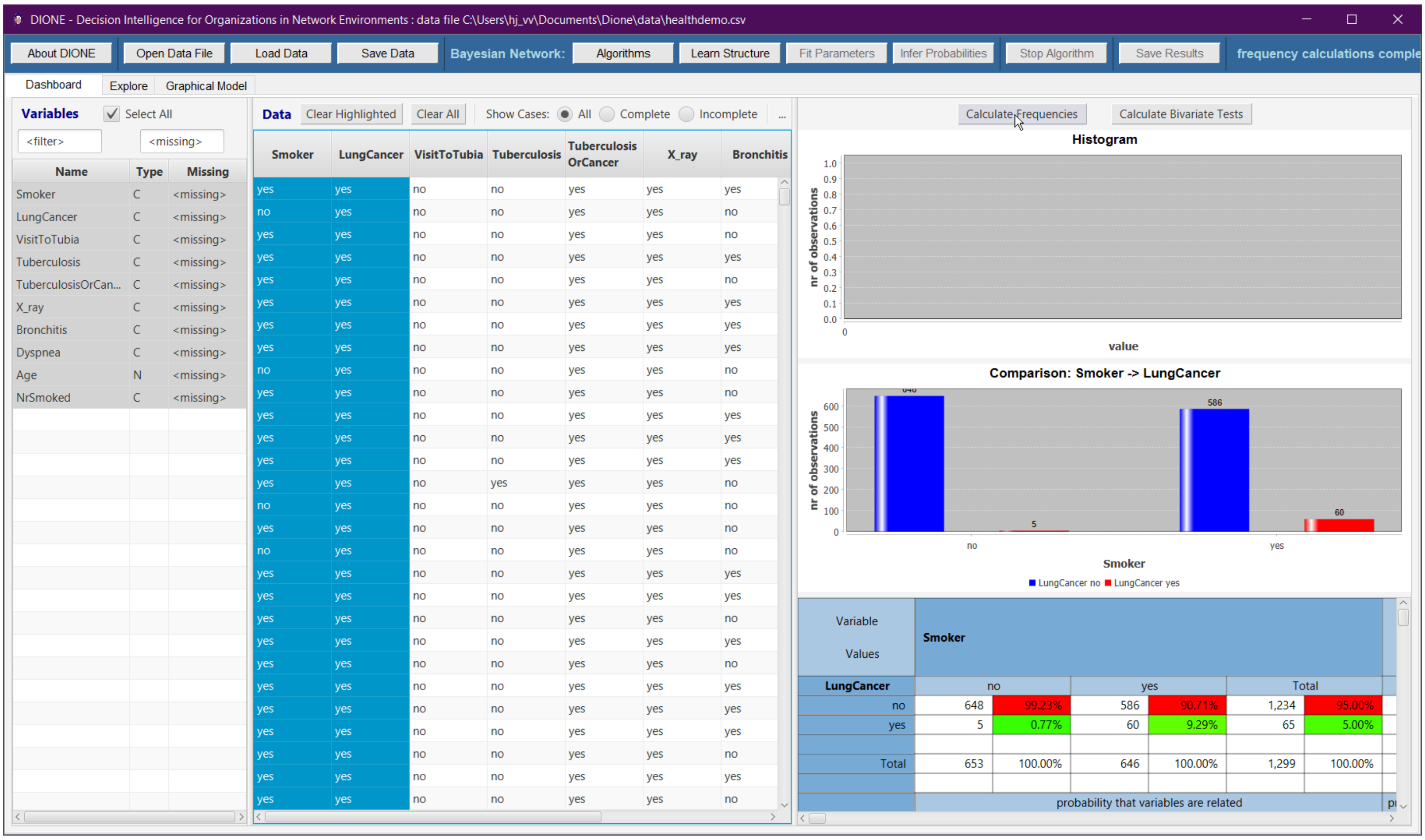
| births_rand... | births_peri... | births_trend0 | births_time... | births |
|---|---|---|---|---|
| NaN | -0.6771947... | NaN | 1 | 26.663 |
| NaN | -2.0829607... | NaN | 2 | 23.598 |
| NaN | 0.86252323... | NaN | 3 | 26.931 |
| NaN | -0.8016786... | NaN | 4 | 24.74 |
| NaN | 0.25165144... | NaN | 5 | 25.806 |
| NaN | -0.1532556... | NaN | 6 | 24.364 |
| -0.9633790... | 1.45604567... | 23.9843333... | 7 | 24.477 |
| -0.9257187... | 1.16459375 | 23.6621250... | 8 | 23.901 |
| -0.9399495... | 0.69161618... | 23.4233333... | 9 | 23.175 |
| -0.7093693... | 0.77524439... | 23.161125 | 10 | 23.227 |
| -0.0824847... | -1.1097652... | 22.8642500... | 11 | 21.672 |
| -0.2983886... | -0.3768197... | 22.5452083... | 12 | 21.87 |
| -0.2373052... | -0.6771947... | 22.3534999... | 13 | 21.439 |
| 0.86325240... | -2.0829607... | 22.3087083... | 14 | 21.089 |
| 0.54389342... | 0.86252323... | 22.3025833... | 15 | 23.709 |
| 0.17588701... | -0.8016786... | 22.2947916... | 16 | 21.669 |
| -0.7931931... | 0.25165144... | 22.2935416... | 17 | 21.752 |
| -1.3913693... | -0.1532556... | 22.305625 | 18 | 20.761 |
| -0.3118790... | 1.45604567... | 22.3348333... | 19 | 23.479 |
| 0.34773958... | 1.16459375 | 22.3116666... | 20 | 23.824 |
| 0.15059214... | 0.69161618... | 22.2627916... | 21 | 23.105 |
| 0.07679727... | 0.77524439... | 22.2579583... | 22 | 23.11 |
| 0.59109855... | -1.1097652... | 22.2776666... | 23 | 21.759 |
| 0.09581971... | -0.3768197... | 22.354 | 24 | 22.073 |
| 0.18381971... | -0.6771947... | 22.430375 | 25 | 21.937 |
| -0.3187059... | -2.0829607... | 22.4366666... | 26 | 20.035 |
| 0.34026842... | 0.86252323... | 22.3872083... | 27 | 23.59 |
| 0.12126201... | -0.8016786... | 22.3524166... | 28 | 21.672 |
| -0.3542347... | 0.25165144... | 22.3245833... | 29 | 22.222 |

| Variable<br>Values | births_timestep | | | | | |
|---|---|---|---|---|---|---|
| births_random0 | [1.0 - 17.7) | | [17.7 - 34.4) | | [34.4 - 51.1) | |
| [-1.8 - -1.5) | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| [-1.5 - -1.1) | 0 | 0.00% | 1 | 5.88% | 0 | 0.00% |
| [-1.1 - -0.8) | 4 | 23.53% | 0 | 0.00% | 0 | 0.00% |
| [-0.8 - -0.4) | 1 | 5.88% | 1 | 5.88% | 5 | 29.41% |
| [-0.4 - -0.1) | 2 | 11.76% | 3 | 17.65% | 2 | 11.76% |
| [-0.1 - 0.2) | 2 | 11.76% | 7 | 41.18% | 6 | 35.29% |

## Two Variables

To compare two variables, select any pair of variables by clicking in the first variable's data column and using `Ctrl-click` anywhere in the second variable's data column: here **Smoker** (`yes/no`) and **LungCancer** (`yes/no`). After pressing button Calculate Frequencies or the `Enter` key, results are shown in the bar chart and in a two-way frequency table, with a preview in the Dashboard tab and a more detailed view in the Explore tab.

In the frequency table, the chi square p-value indicates the significance of the relation between the variables (the lower the p-value, the more significant). The risk ratio is also calculated with corresponding confidence interval and p-values. The risk ratio (or relative risk) compares the risks for two groups by dividing the incidence in the exposed group by the incidence in the unexposed group. Thus, in the example, the percentage of smokers with lung cancer is 12.13 times the percentage of non-smokers with lung cancer. The risk ratio p-values show that this risk ratio value is significant, and the 95 % confidence interval means that the value is between 4.90 and 30.01 with 95 % probability.

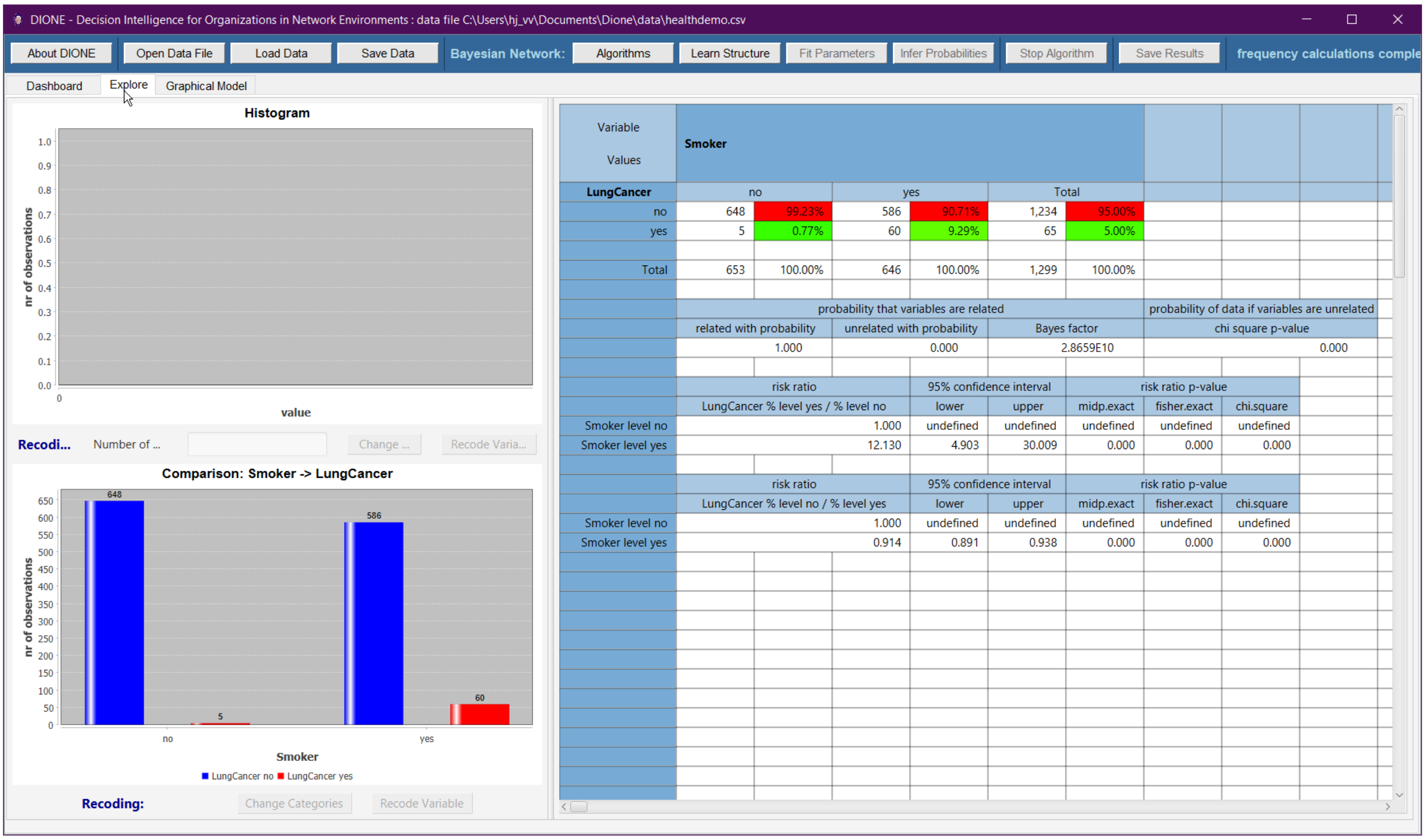


DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemo.csv

| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | C | <missing> |

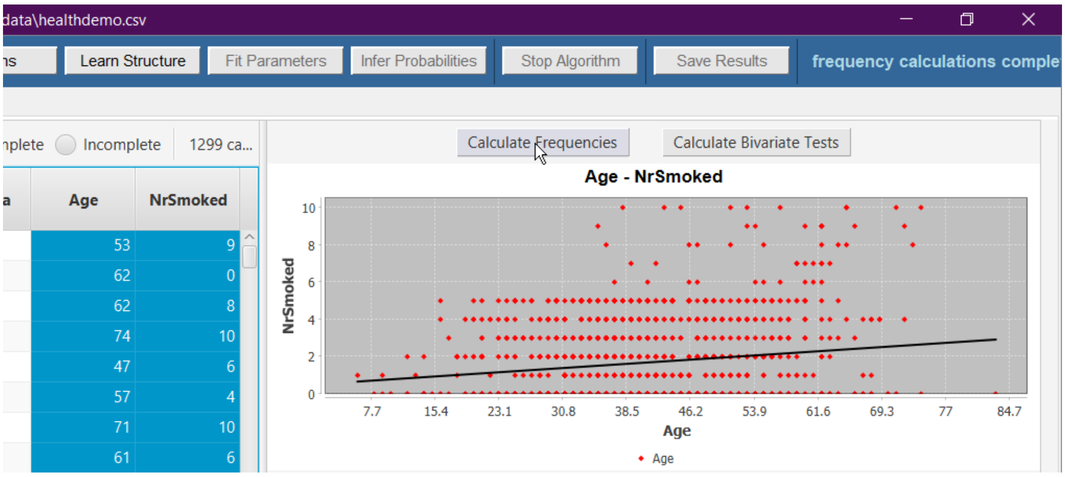
| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis |
|---|---|---|---|---|---|---|
| yes | yes | no | no | yes | yes | yes |
| no | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes |
| no | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | yes | yes | yes | no |
| no | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | no |
| no | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | no |
| yes | yes | no | no | yes | yes | yes |
| yes | yes | no | no | yes | yes | no |

| Variable<br>Values | Smoker | | | | | |
|---|---|---|---|---|---|---|
| **LungCancer** | no | | yes | | Total | |
| no | 648 | 99.23% | 586 | 90.71% | 1,234 | 95.00% |
| yes | 5 | 0.77% | 60 | 9.29% | 65 | 5.00% |
| | | | | | | |
| Total | 653 | 100.00% | 646 | 100.00% | 1,299 | 100.00% |
| | | | | | | |
| | probability that variables are related | | | | | |



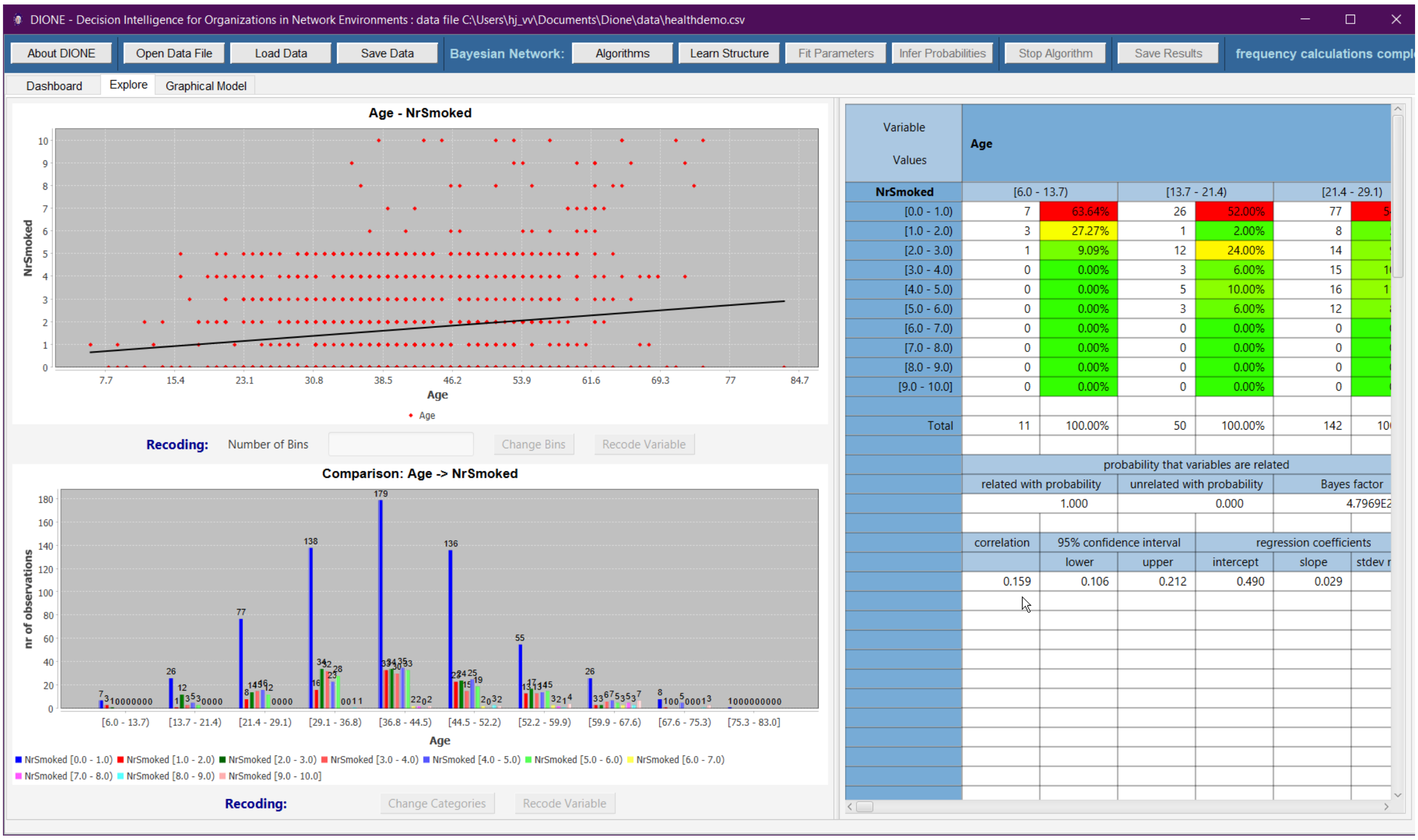
| Variable<br>Values | Smoker | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| LungCancer | no | | yes | | Total | | | | |
| no | 648 | 99.23% | 586 | 90.71% | 1,234 | 95.00% | | | |
| yes | 5 | 0.77% | 60 | 9.29% | 65 | 5.00% | | | |
| | | | | | | | | | |
| Total | 653 | 100.00% | 646 | 100.00% | 1,299 | 100.00% | | | |
| | | | | | | | | | |
| | probability that variables are related | | | | | | probability of data if variables are unrelated | | |
| | related with probability | | unrelated with probability | | Bayes factor | | chi square p-value | | |
| | 1.000 | | 0.000 | | 2.8659E10 | | 0.000 | | |
| | | | | | | | | | |
| | risk ratio | | | 95% confidence interval | | risk ratio p-value | | | |
| | LungCancer % level yes / % level no | | | lower | upper | midp.exact | fisher.exact | chi.square | |
| Smoker level no | 1.000 | | | undefined | undefined | undefined | undefined | undefined | |
| Smoker level yes | 12.130 | | | 4.903 | 30.009 | 0.000 | 0.000 | 0.000 | |
| | | | | | | | | | |
| | risk ratio | | | 95% confidence interval | | risk ratio p-value | | | |
| | LungCancer % level no / % level yes | | | lower | upper | midp.exact | fisher.exact | chi.square | |
| Smoker level no | 1.000 | | | undefined | undefined | undefined | undefined | undefined | |
| Smoker level yes | 0.914 | | | 0.891 | 0.938 | 0.000 | 0.000 | 0.000 | |

Frequencies can also be calculated for two numerical variables, for example **Age** and **NrSmoked**, after setting **NrSmoked** to **Type** numerical (N) in the **Variables** table. As shown here, they are compared in a scatter plot, including a linear line of best fit (regression line):



| a | Age | NrSmoked |
|---|---|---|
| | 53 | 9 |
| | 62 | 0 |
| | 62 | 8 |
| | 74 | 10 |
| | 47 | 6 |
| | 57 | 4 |
| | 71 | 10 |
| | 61 | 6 |

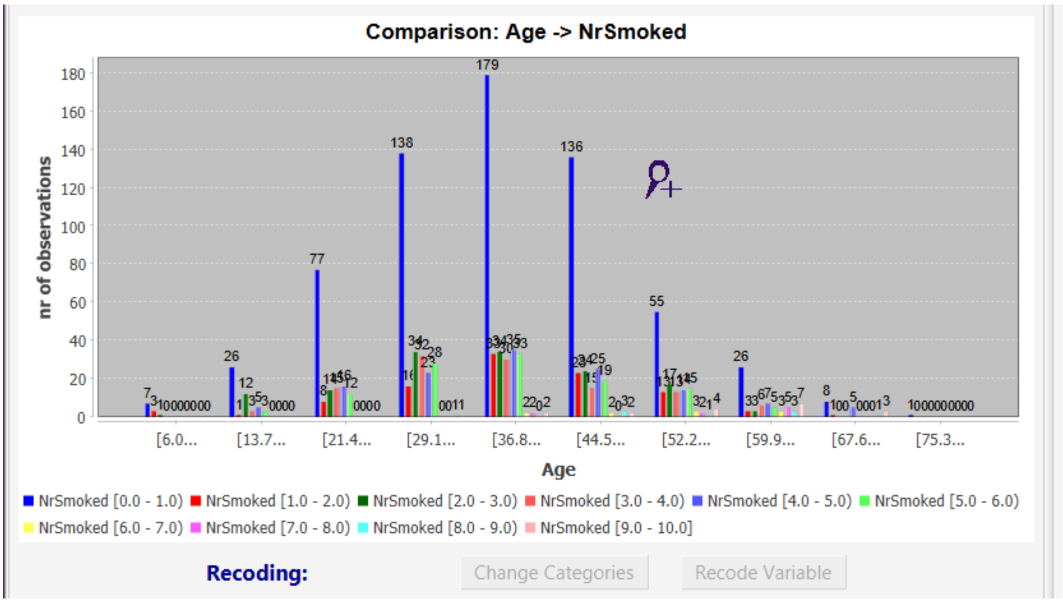
Their correlation coefficient is shown in the results table after the frequency data. Correlation coefficient 1 means perfect correlation and 0 means no correlation, so in the example there is a weak correlation, as is also clear from the scatter plot. Regression coefficients are also shown.

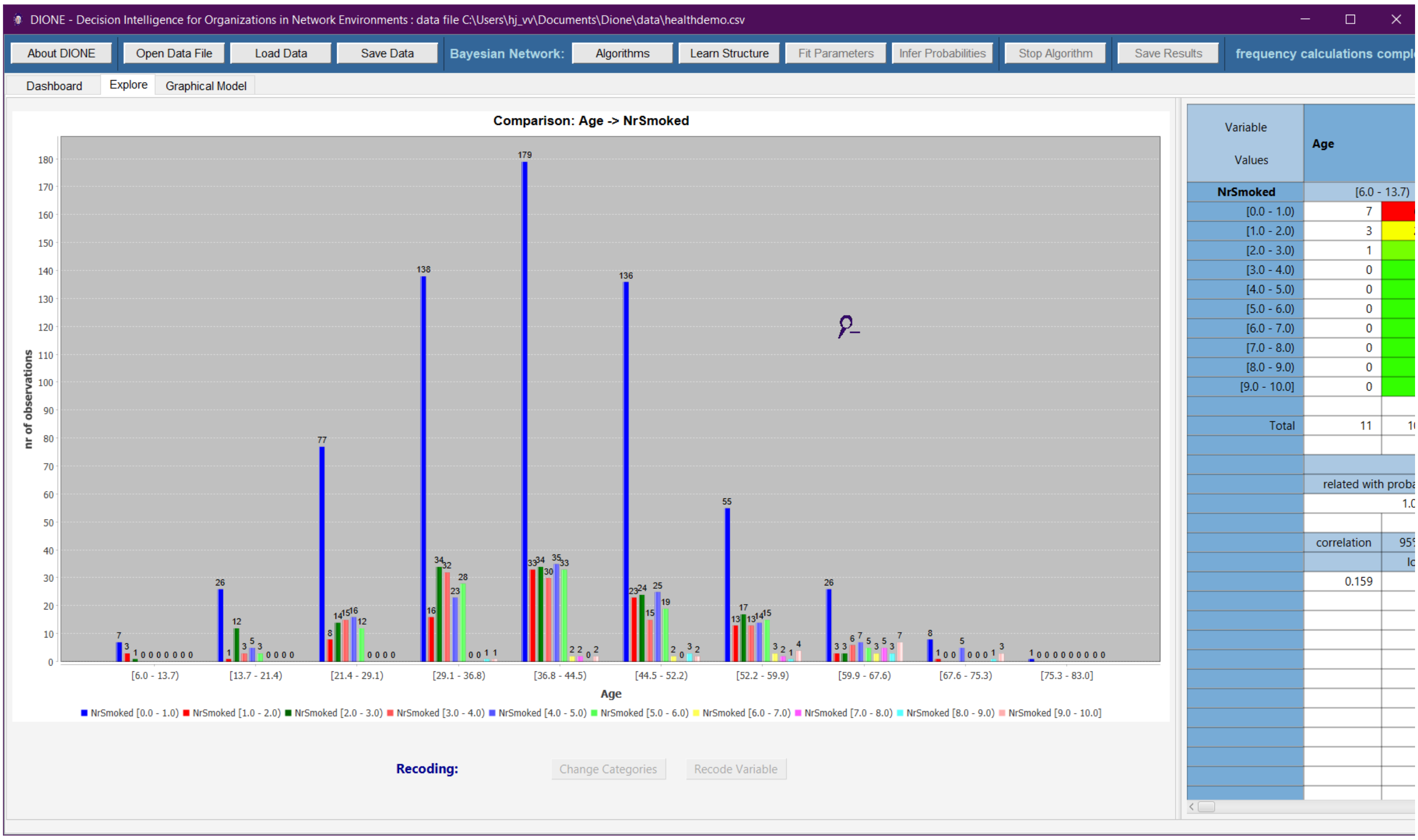


| Variable<br>Values | Age | | | | | |
|---|---|---|---|---|---|---|
| NrSmoked | [6.0 - 13.7) | | [13.7 - 21.4) | | [21.4 - 29.1) | |
| [0.0 - 1.0) | 7 | 63.64% | 26 | 52.00% | 77 | 5 |
| [1.0 - 2.0) | 3 | 27.27% | 1 | 2.00% | 8 | |
| [2.0 - 3.0) | 1 | 9.09% | 12 | 24.00% | 14 | |
| [3.0 - 4.0) | 0 | 0.00% | 3 | 6.00% | 15 | 1 |
| [4.0 - 5.0) | 0 | 0.00% | 5 | 10.00% | 16 | 1 |
| [5.0 - 6.0) | 0 | 0.00% | 3 | 6.00% | 12 | |
| [6.0 - 7.0) | 0 | 0.00% | 0 | 0.00% | 0 | |
| [7.0 - 8.0) | 0 | 0.00% | 0 | 0.00% | 0 | |
| [8.0 - 9.0) | 0 | 0.00% | 0 | 0.00% | 0 | |
| [9.0 - 10.0] | 0 | 0.00% | 0 | 0.00% | 0 | |
| | | | | | | |
| Total | 11 | 100.00% | 50 | 100.00% | 142 | 10 |
| | | | | | | |
| | probability that variables are related | | | | | |
| | related with probability | | unrelated with probability | | Bayes factor | |
| | 1.000 | | 0.000 | | 4.7969E2 | |
| | | | | | | |
| | correlation | 95% confidence interval | | regression coefficients | | |
| | | lower | upper | intercept | slope | stdev r |
| | 0.159 | 0.106 | 0.212 | 0.490 | 0.029 | |

## Inspect Results

A preview of results is shown in the Dashboard tab and a more detailed view in the Explore tab. Sometimes a variable has many levels and/or long texts as values, so the chart become hard to read. In the Explore tab you can zoom in on a chart by hovering over it with the mouse to show an hourglass with a + sign, then clicking to zoom in:



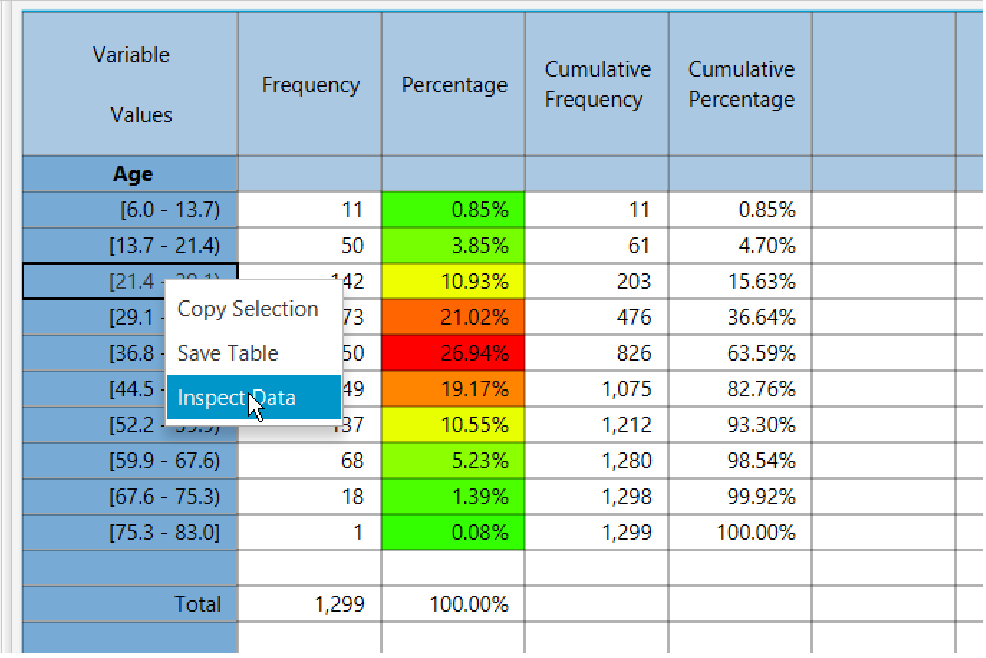


frequency calculations compl

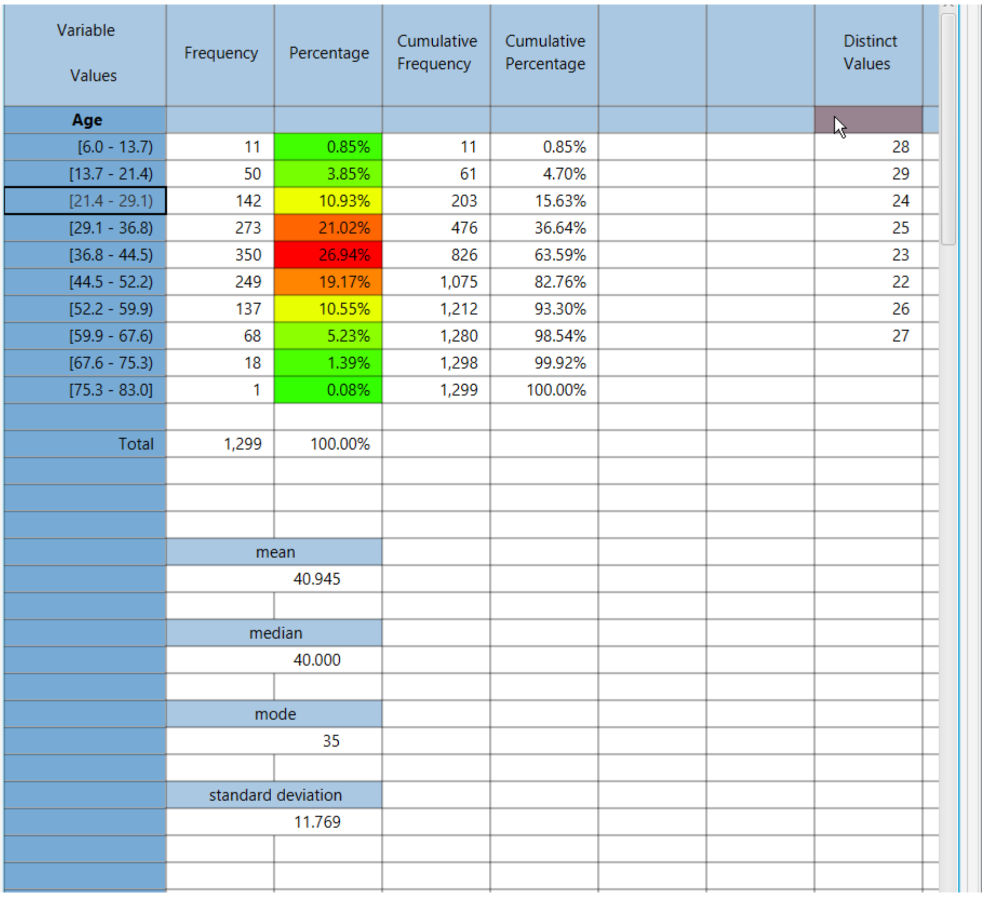
| Variable<br>Values | Age | |
|---|---|---|
| NrSmoked | [6.0 - 13.7) | |
| [0.0 - 1.0) | 7 | 🟥 |
| [1.0 - 2.0) | 3 | 🟨 |
| [2.0 - 3.0) | 1 | 🟩 |
| [3.0 - 4.0) | 0 | 🟩 |
| [4.0 - 5.0) | 0 | 🟩 |
| [5.0 - 6.0) | 0 | 🟩 |
| [6.0 - 7.0) | 0 | 🟩 |
| [7.0 - 8.0) | 0 | 🟩 |
| [8.0 - 9.0) | 0 | 🟩 |
| [9.0 - 10.0] | 0 | 🟩 |
| | | |
| Total | 11 | 1( |
| | | |
| | related with proba | |
| | 1.( | |
| | | |
| | correlation | 95' |
| | | lc |
| | 0.159 | |

The hourglass now has a – sign to zoom out again.

At the right-hand side of the Explore tab, there is a table with detailed frequency data and other analysis results. For numerical variables, frequencies of ranges of values are shown. If desired, you can get more details about individual values of a variable by right-clicking on one of the value ranges in the left-hand column and choosing menu item Inspect Data to see the distinct values of that variable:



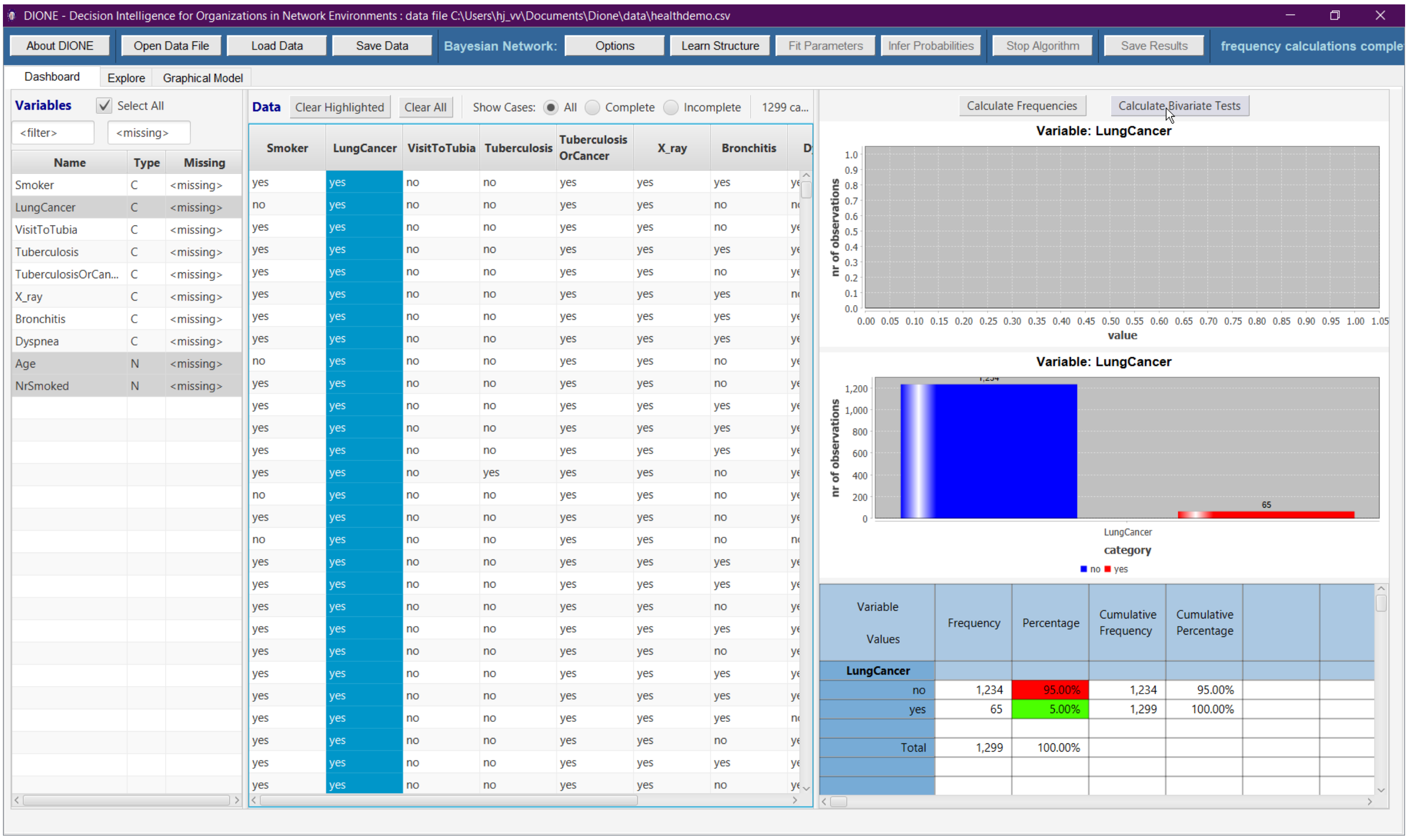
| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| Age | | | | |
| [6.0 - 13.7) | 11 | 0.85% | 11 | 0.85% |
| [13.7 - 21.4) | 50 | 3.85% | 61 | 4.70% |
| [21.4 - | 142 | 10.93% | 203 | 15.63% |
| [29.1 - | 73 | 21.02% | 476 | 36.64% |
| [36.8 - | 50 | 26.94% | 826 | 63.59% |
| [44.5 - | 49 | 19.17% | 1,075 | 82.76% |
| [52.2 - | 137 | 10.55% | 1,212 | 93.30% |
| [59.9 - 67.6) | 68 | 5.23% | 1,280 | 98.54% |
| [67.6 - 75.3) | 18 | 1.39% | 1,298 | 99.92% |
| [75.3 - 83.0] | 1 | 0.08% | 1,299 | 100.00% |
| | | | | |
| Total | 1,299 | 100.00% | | |



| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage | | | Distinct Values |
|---|---|---|---|---|---|---|---|
| **Age** | | | | | | | |
| [6.0 - 13.7) | 11 | 0.85% | 11 | 0.85% | | | 28 |
| [13.7 - 21.4) | 50 | 3.85% | 61 | 4.70% | | | 29 |
| [21.4 - 29.1) | 142 | 10.93% | 203 | 15.63% | | | 24 |
| [29.1 - 36.8) | 273 | 21.02% | 476 | 36.64% | | | 25 |
| [36.8 - 44.5) | 350 | 26.94% | 826 | 63.59% | | | 23 |
| [44.5 - 52.2) | 249 | 19.17% | 1,075 | 82.76% | | | 22 |
| [52.2 - 59.9) | 137 | 10.55% | 1,212 | 93.30% | | | 26 |
| [59.9 - 67.6) | 68 | 5.23% | 1,280 | 98.54% | | | 27 |
| [67.6 - 75.3) | 18 | 1.39% | 1,298 | 99.92% | | | |
| [75.3 - 83.0] | 1 | 0.08% | 1,299 | 100.00% | | | |
| | | | | | | | |
| Total | 1,299 | 100.00% | | | | | |
| | mean | | | | | | |
| | 40.945 | | | | | | |
| | median | | | | | | |
| | 40.000 | | | | | | |
| | mode | | | | | | |
| | 35 | | | | | | |
| | standard deviation | | | | | | |
| | 11.769 | | | | | | |

## All Variables

To obtain bivariate comparisons of one variable with other variables, start by selecting, in the Dashboard tab, one variable as the outcome variable of interest. In this example the outcome variable is **LungCancer**, with which all other loaded variables will be compared:



DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemo.csv

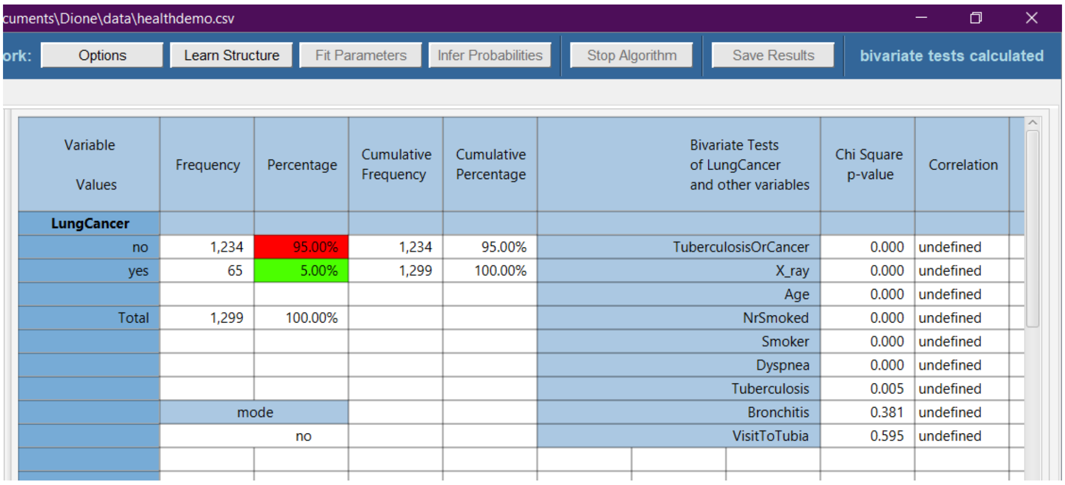
| Name | Type | Missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | N | <missing> |

| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis | D |
|---|---|---|---|---|---|---|---|
| yes | yes | no | no | yes | yes | yes | ye |
| no | yes | no | no | yes | yes | no | n |
| yes | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | yes | n |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| no | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | yes | yes | yes | no | ye |
| no | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | no | ye |
| no | yes | no | no | yes | yes | no | n |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | yes | n |
| yes | yes | no | no | yes | yes | no | ye |
| yes | yes | no | no | yes | yes | yes | ye |
| yes | yes | no | no | yes | yes | no | ye |

| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage | | |
|---|---|---|---|---|---|---|
| **LungCancer** | | | | | | |
| no | 1,234 | 95.00% | 1,234 | 95.00% | | |
| yes | 65 | 5.00% | 1,299 | 100.00% | | |
| | | | | | | |
| Total | 1,299 | 100.00% | | | | |

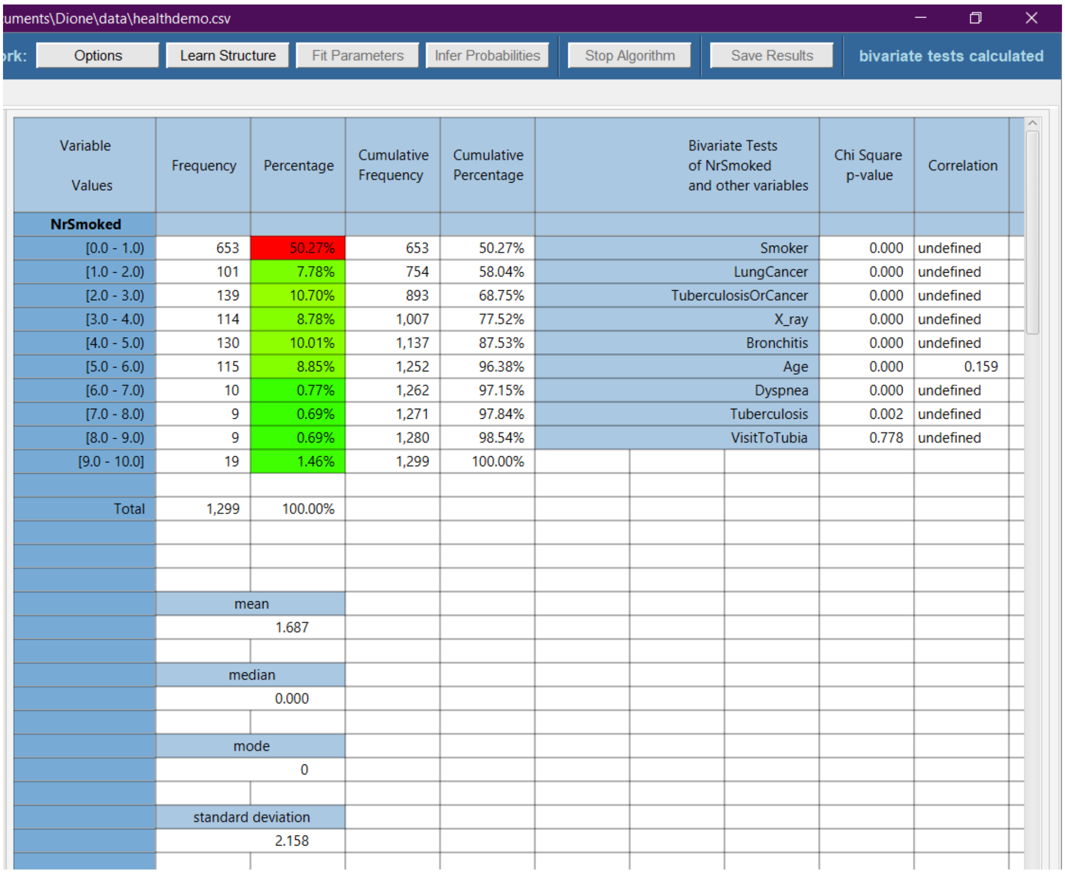
Press button Calculate Bivariate Tests above the plots (histogram and bar chart) at the right.

Go to the Explore tab and scroll to the right in the results table to see the p-values of chi square tests of the selected outcome variable compared with all other loaded variables, ordered from most to least significant association. This allows you to see quickly which variables have the most significant association with the selected outcome variable and so are likely to be of interest for further analysis. In the example all variables except **Bronchitis** and **VisitToTubia** have significant associations at the 1 % (so also of course at the 5 %) significance level with **LungCancer** ($p < 0.01$ and $p < 0.05$).



| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage | Bivariate Tests of LungCancer and other variables | Chi Square p-value | Correlation |
|---|---|---|---|---|---|---|---|
| **LungCancer** | | | | | | | |
| no | 1,234 | 95.00% | 1,234 | 95.00% | TuberculosisOrCancer | 0.000 | undefined |
| yes | 65 | 5.00% | 1,299 | 100.00% | X_ray | 0.000 | undefined |
| | | | | | Age | 0.000 | undefined |
| Total | 1,299 | 100.00% | | | NrSmoked | 0.000 | undefined |
| | | | | | Smoker | 0.000 | undefined |
| | | | | | Dyspnea | 0.000 | undefined |
| | | | | | Tuberculosis | 0.005 | undefined |
| | mode | | | | Bronchitis | 0.381 | undefined |
| | | no | | | VisitToTubia | 0.595 | undefined |

For numerical data correlations of all variables with the outcome variable are also calculated. In this case results are ordered first according to increasing p-value of the chi square tests, and variables with approximately equal p-values (i.e. with difference in p-values less than 0.01) are ordered according to decreasing correlation (i.e. higher to lower absolute value of the correlation coefficient). This example shows results for variable **NrSmoked**:

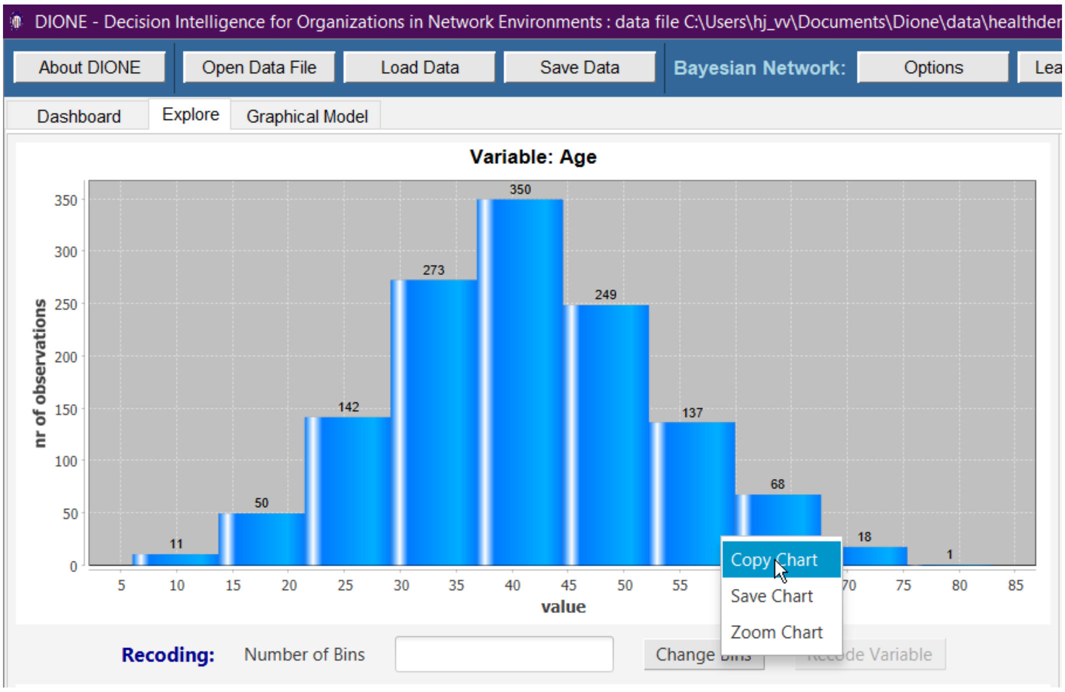


| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage | Bivariate Tests of NrSmoked and other variables | Chi Square p-value | Correlation |
|---|---|---|---|---|---|---|---|
| **NrSmoked** | | | | | | | |
| [0.0 - 1.0) | 653 | 50.27% | 653 | 50.27% | Smoker | 0.000 | undefined |
| [1.0 - 2.0) | 101 | 7.78% | 754 | 58.04% | LungCancer | 0.000 | undefined |
| [2.0 - 3.0) | 139 | 10.70% | 893 | 68.75% | TuberculosisOrCancer | 0.000 | undefined |
| [3.0 - 4.0) | 114 | 8.78% | 1,007 | 77.52% | X_ray | 0.000 | undefined |
| [4.0 - 5.0) | 130 | 10.01% | 1,137 | 87.53% | Bronchitis | 0.000 | undefined |
| [5.0 - 6.0) | 115 | 8.85% | 1,252 | 96.38% | Age | 0.000 | 0.159 |
| [6.0 - 7.0) | 10 | 0.77% | 1,262 | 97.15% | Dyspnea | 0.000 | undefined |
| [7.0 - 8.0) | 9 | 0.69% | 1,271 | 97.84% | Tuberculosis | 0.002 | undefined |
| [8.0 - 9.0) | 9 | 0.69% | 1,280 | 98.54% | VisitToTubia | 0.778 | undefined |
| [9.0 - 10.0] | 19 | 1.46% | 1,299 | 100.00% | | | |
| | | | | | | | |
| Total | 1,299 | 100.00% | | | | | |
| | mean | | | | | | |
| | | 1.687 | | | | | |
| | median | | | | | | |
| | | 0.000 | | | | | |
| | mode | | | | | | |
| | | 0 | | | | | |
| | standard deviation | | | | | | |
| | | 2.158 | | | | | |

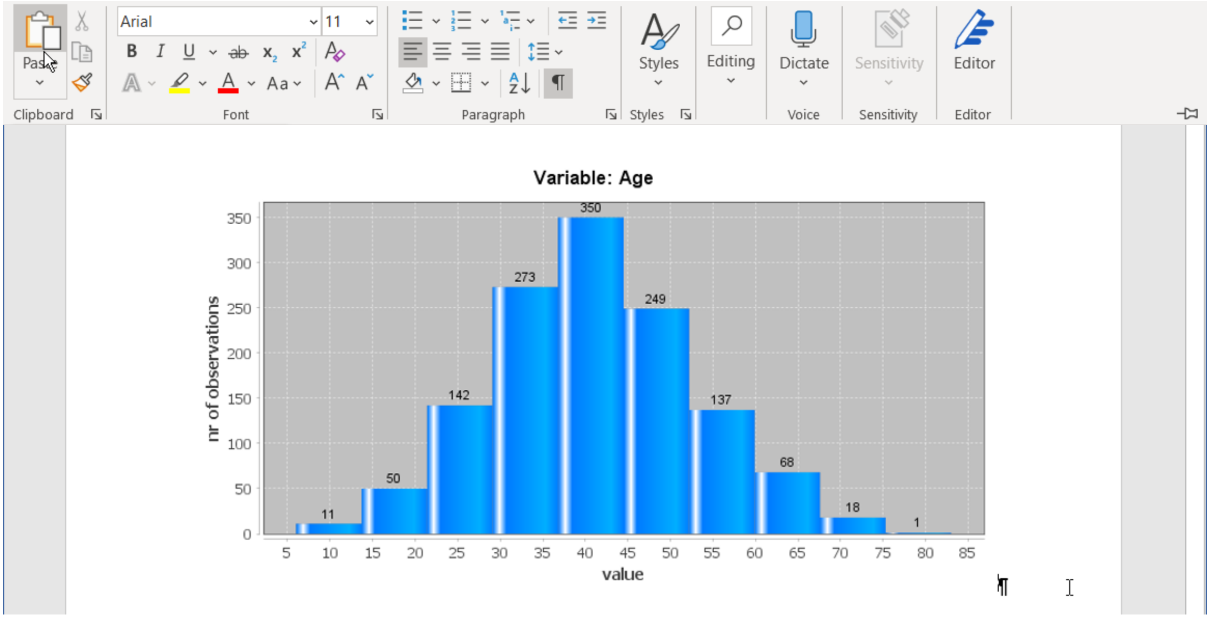
## Copy and Save Results

You can keep your results by copying charts and tables to another application, for example to insert them in a word processing document, or by saving them to a file.

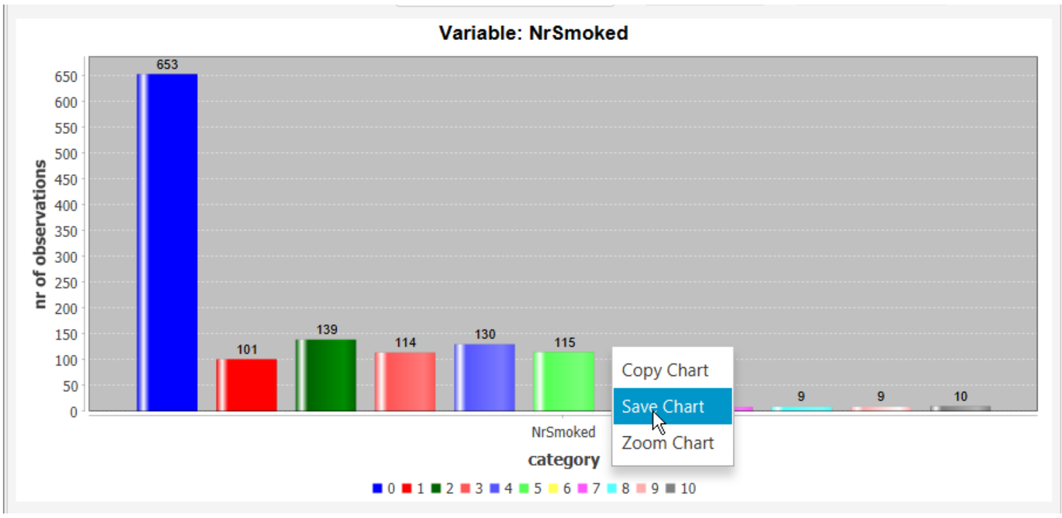
To copy a bar chart or a histogram, right-click on it and select Copy Chart from the pop-up menu to put a picture of the chart on your computer's clipboard:

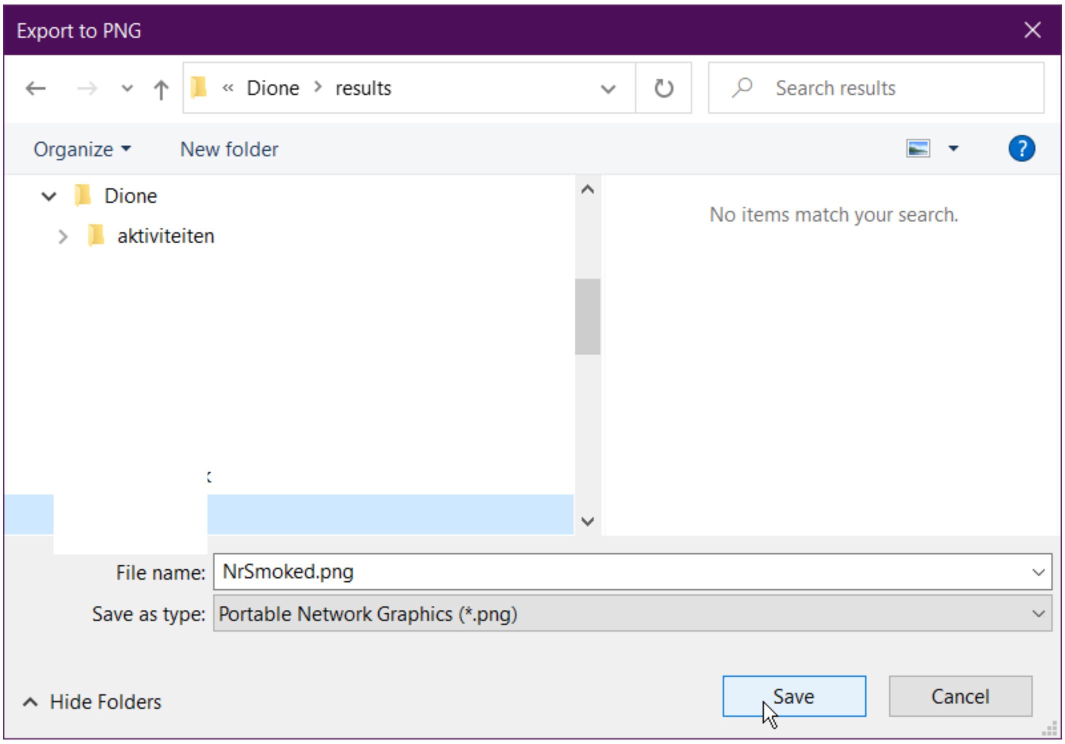


You can now paste the picture in a Word document, for example:



To save a chart to a file, right-click on it and select Save Chart from the pop-up menu to open a save dialog:





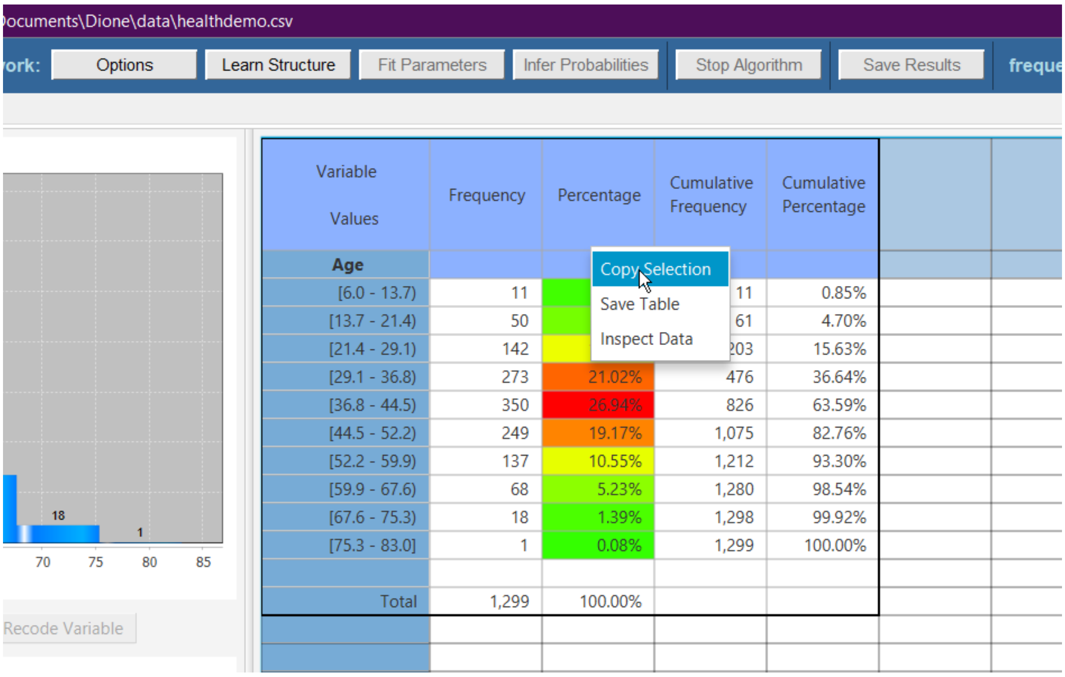
File name: NrSmoked.png

Save as type: Portable Network Graphics (*.png)

You now have a Portable Network Graphics (PNG) picture file that you van open in any image processing program or insert in a document.

There is another option, Zoom Chart , in the pop-up menu that you can use to zoom in or out a chart as with the zoom hourglass, if you prefer.

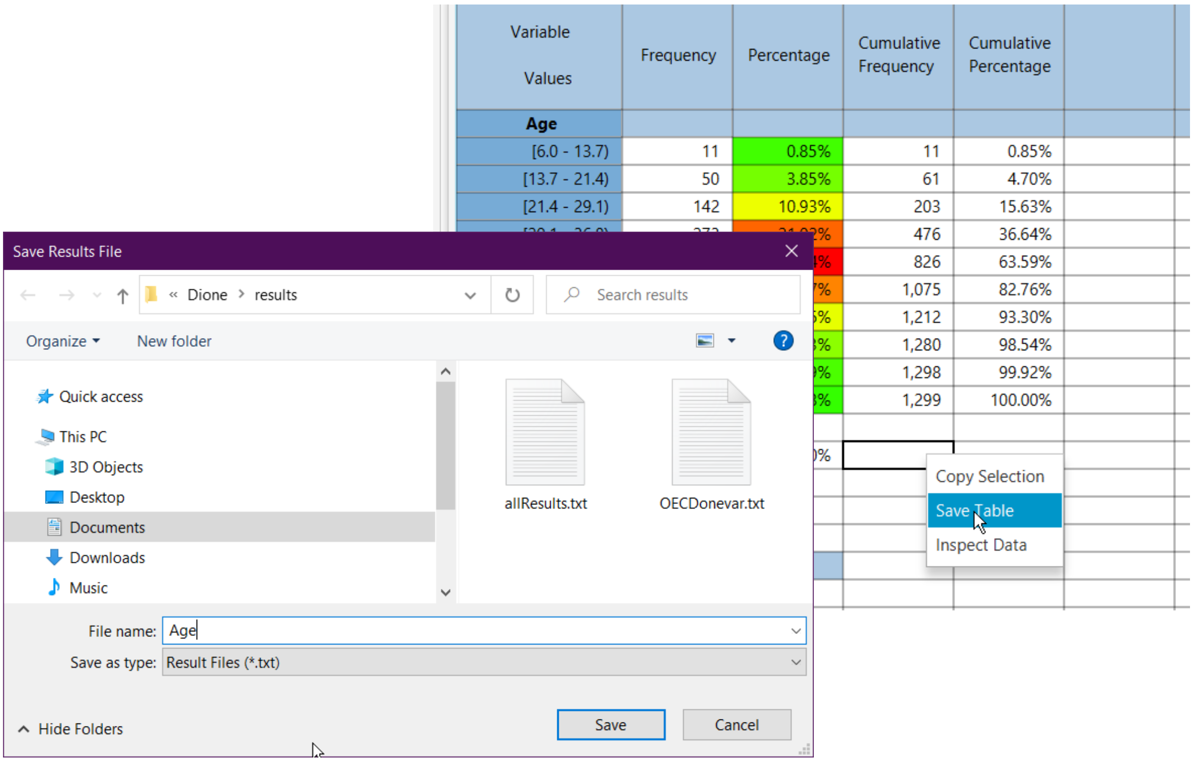
You can copy results in the frequency table by right-clicking on a selection of data you want to copy and choosing Copy Selection in the pop-up menu:



| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| **Age** | | | | |
| [6.0 - 13.7) | 11 | | 11 | 0.85% |
| [13.7 - 21.4) | 50 | | 61 | 4.70% |
| [21.4 - 29.1) | 142 | | 203 | 15.63% |
| [29.1 - 36.8) | 273 | 21.02% | 476 | 36.64% |
| [36.8 - 44.5) | 350 | 26.94% | 826 | 63.59% |
| [44.5 - 52.2) | 249 | 19.17% | 1,075 | 82.76% |
| [52.2 - 59.9) | 137 | 10.55% | 1,212 | 93.30% |
| [59.9 - 67.6) | 68 | 5.23% | 1,280 | 98.54% |
| [67.6 - 75.3) | 18 | 1.39% | 1,298 | 99.92% |
| [75.3 - 83.0] | 1 | 0.08% | 1,299 | 100.00% |
| | | | | |
| Total | 1,299 | 100.00% | | |

Copied data can be pasted in a Word document, for example, in the same way as described for a chart.

You can save the whole table in a text file by right-clicking anywhere in the table and choosing Save Table in the pop-up menu to bring up a save dialog:



| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| **Age** | | | | |
| [6.0 - 13.7) | 11 | 0.85% | 11 | 0.85% |
| [13.7 - 21.4) | 50 | 3.85% | 61 | 4.70% |
| [21.4 - 29.1) | 142 | 10.93% | 203 | 15.63% |
| | | | 476 | 36.64% |
| | | | 826 | 63.59% |
| | | | 1,075 | 82.76% |
| | | | 1,212 | 93.30% |
| | | | 1,280 | 98.54% |
| | | | 1,298 | 99.92% |
| | | | 1,299 | 100.00% |

File name: Age

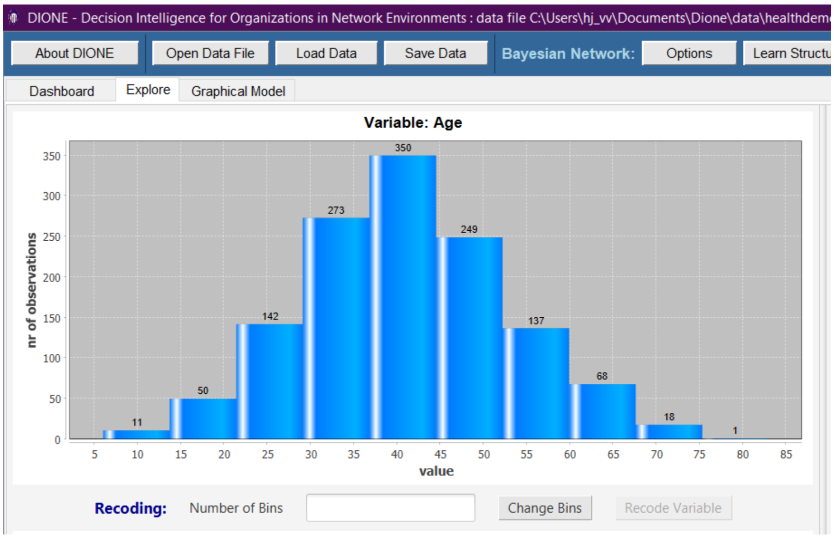
Save as type: Result Files (*.txt)

# Recode Data

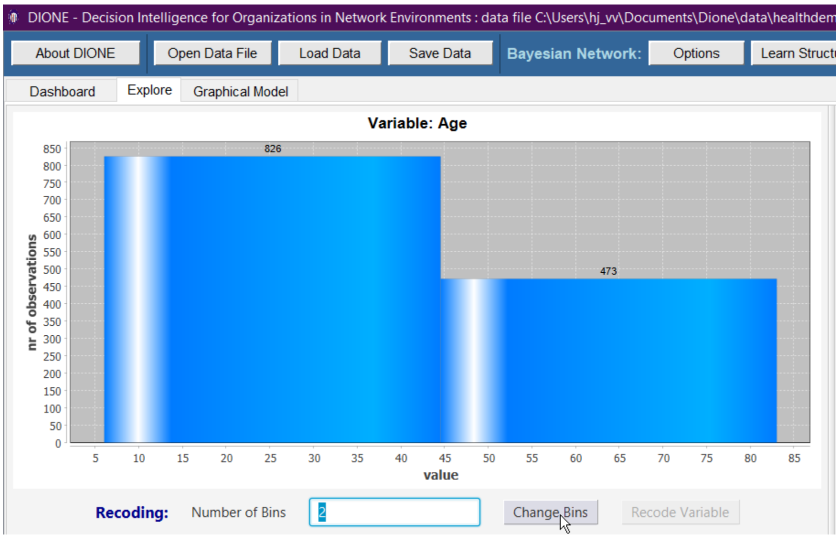
When a variable has numerical data, sometimes more meaningful results can be obtained by recoding it to categorical data. An example is variable **Age** in the second example of the **Explore Data** section. You can recode such variables to categorical variables in the Explore tab. You can also recode categorical variables to new categorical variables with different categories, typically to reduce the number of categories or to give categories meaningful names. Recoding a variable does not replace the existing variable but creates a new variable with a new name.

## Numerical to Categorical

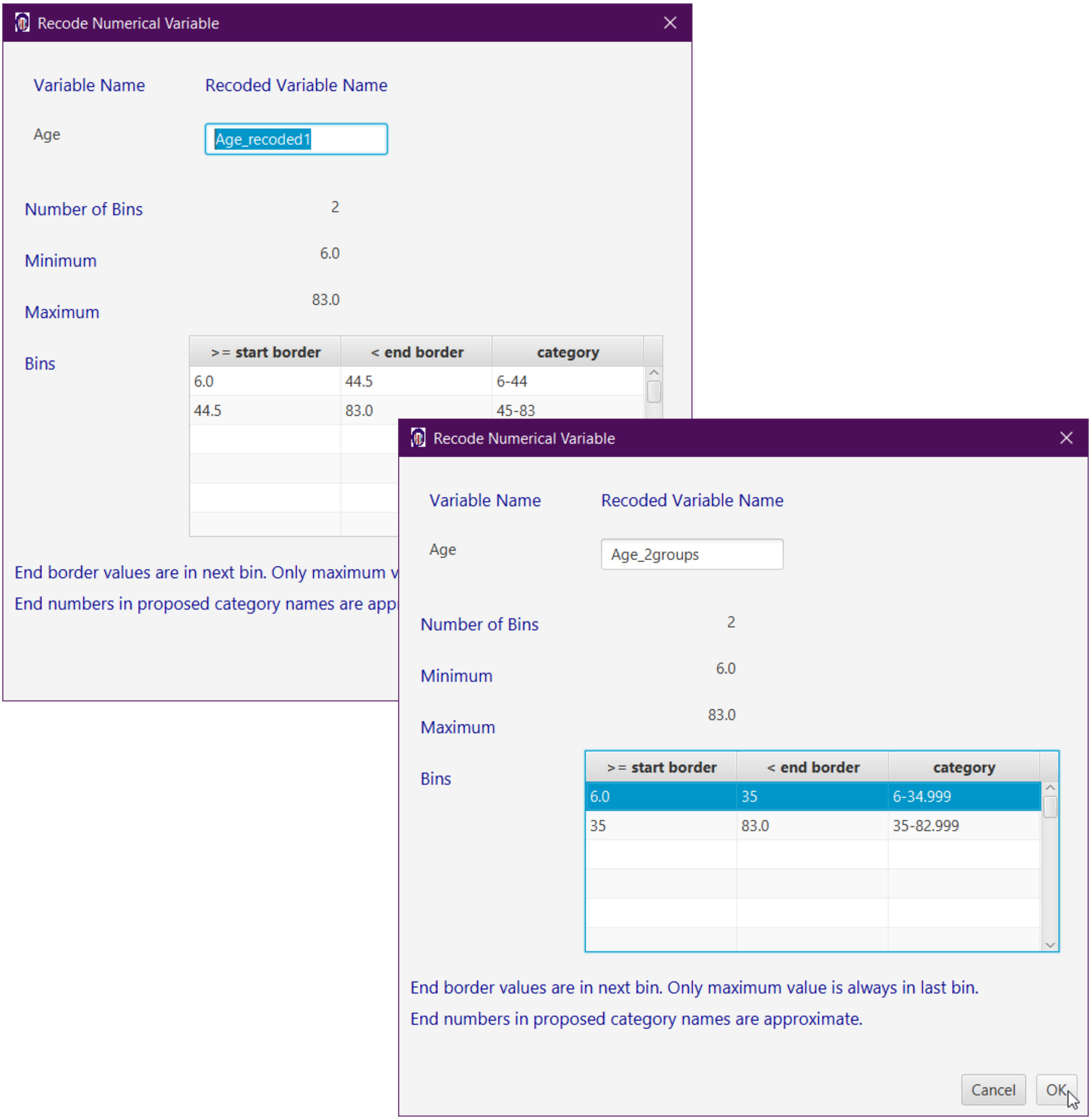
For example, we would like to recode the variable **Age**, which has values from 6 to 83 years old, to a variable with two categories (age groups). When **Age** is selected in the Dashboard tab, the Explore tab shows this histogram:



Type the value 2 in the Number of Bins text field and press `Enter` to organize the **Age** data in two categories. The cut-off point between the two categories is automatically made:



Category borders can be changed by pressing button Change Bins to open the recoding dialog, in which the recoded variable name and its category names can also be changed:



| Variable Name | Recoded Variable Name |
|---|---|
| Age | Age_recoded1 |

Number of Bins 2

Minimum 6.0

Maximum 83.0

Bins

| >= start border | < end border | category |
|---|---|---|
| 6.0 | 44.5 | 6-44 |
| 44.5 | 83.0 | 45-83 |

End border values are in next bin. Only maximum v

End numbers in proposed category names are app

| Variable Name | Recoded Variable Name |
|---|---|
| Age | Age_2groups |

Number of Bins 2

Minimum 6.0

Maximum 83.0

Bins

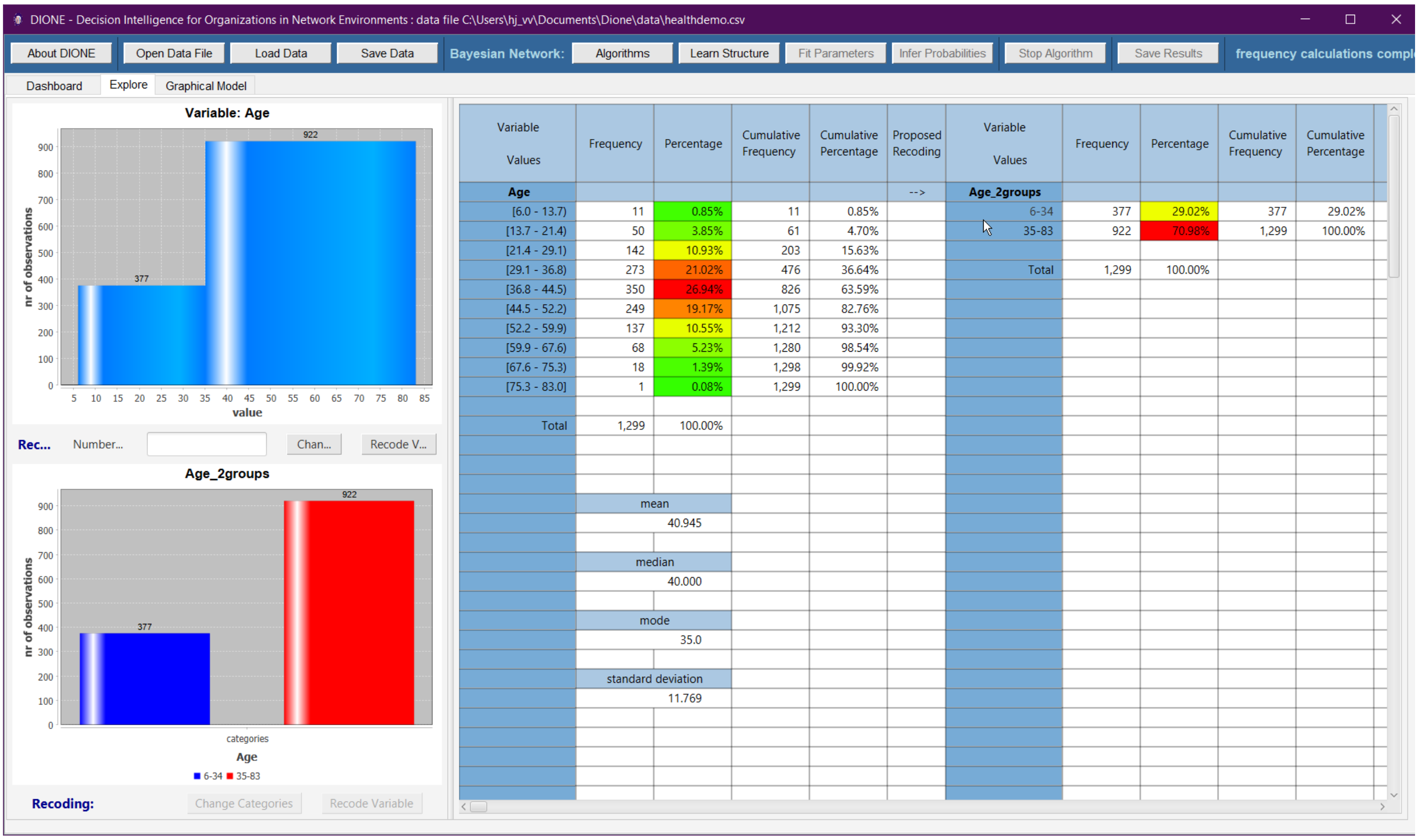
| >= start border | < end border | category |
|---|---|---|
| 6.0 | 35 | 6-34.999 |
| 35 | 83.0 | 35-82.999 |

End border values are in next bin. Only maximum value is always in last bin.

End numbers in proposed category names are approximate.

In the Recoded Variable Name field a name is proposed for the new variable, which you can accept or change to a name you prefer. The new name cannot be the name of an existing variable; if the name is already in use, an error message is shown. In the < end border column of the Bins table, type the desired cut-off value(s) (the cut-off value itself will be in the next higher bin), then press `OK`. Values in the table are easily edited by using arrow keys to change rows: for example, immediately after editing a cell, use the `Down Arrow` key to edit the cell below.

The proposed recoding preview (histogram, bar chart and frequency table) will now look like this:



| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage | Proposed Recoding | Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|---|---|---|---|---|---|
| Age | | | | | --> | Age_2groups | | | | |
| [6.0 - 13.7) | 11 | 0.85% | 11 | 0.85% | | 6-34 | 377 | 29.02% | 377 | 29.02% |
| [13.7 - 21.4) | 50 | 3.85% | 61 | 4.70% | | 35-83 | 922 | 70.98% | 1,299 | 100.00% |
| [21.4 - 29.1) | 142 | 10.93% | 203 | 15.63% | | | | | | |
| [29.1 - 36.8) | 273 | 21.02% | 476 | 36.64% | | Total | 1,299 | 100.00% | | |
| [36.8 - 44.5) | 350 | 26.94% | 826 | 63.59% | | | | | | |
| [44.5 - 52.2) | 249 | 19.17% | 1,075 | 82.76% | | | | | | |
| [52.2 - 59.9) | 137 | 10.55% | 1,212 | 93.30% | | | | | | |
| [59.9 - 67.6) | 68 | 5.23% | 1,280 | 98.54% | | | | | | |
| [67.6 - 75.3) | 18 | 1.39% | 1,298 | 99.92% | | | | | | |
| [75.3 - 83.0] | 1 | 0.08% | 1,299 | 100.00% | | | | | | |
| | | | | | | | | | | |
| Total | 1,299 | 100.00% | | | | | | | | |

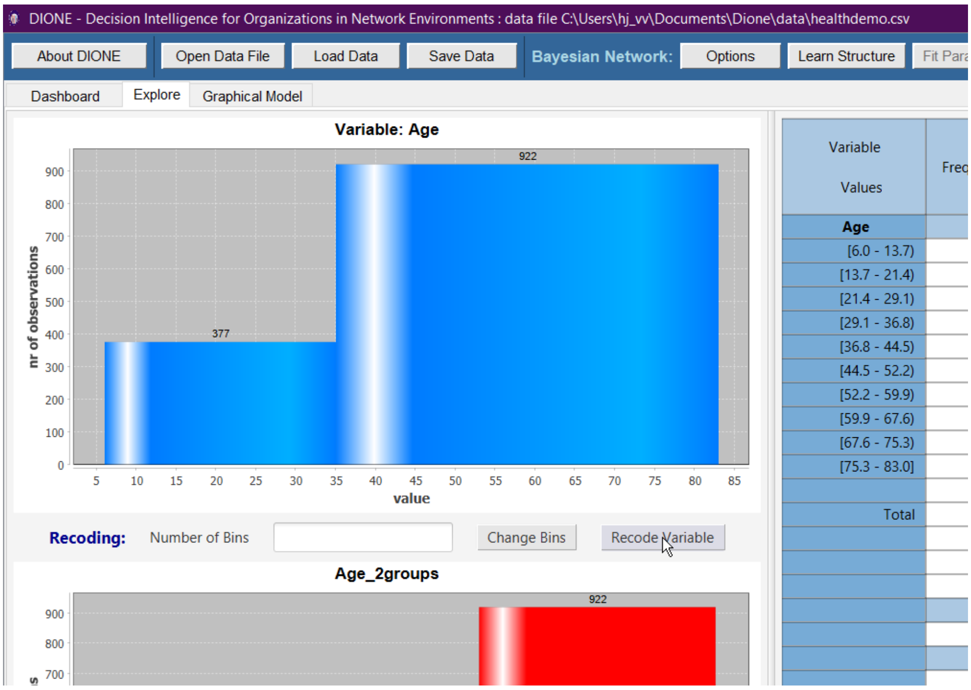
mean
40.945

median
40.000

mode
35.0

standard deviation
11.769

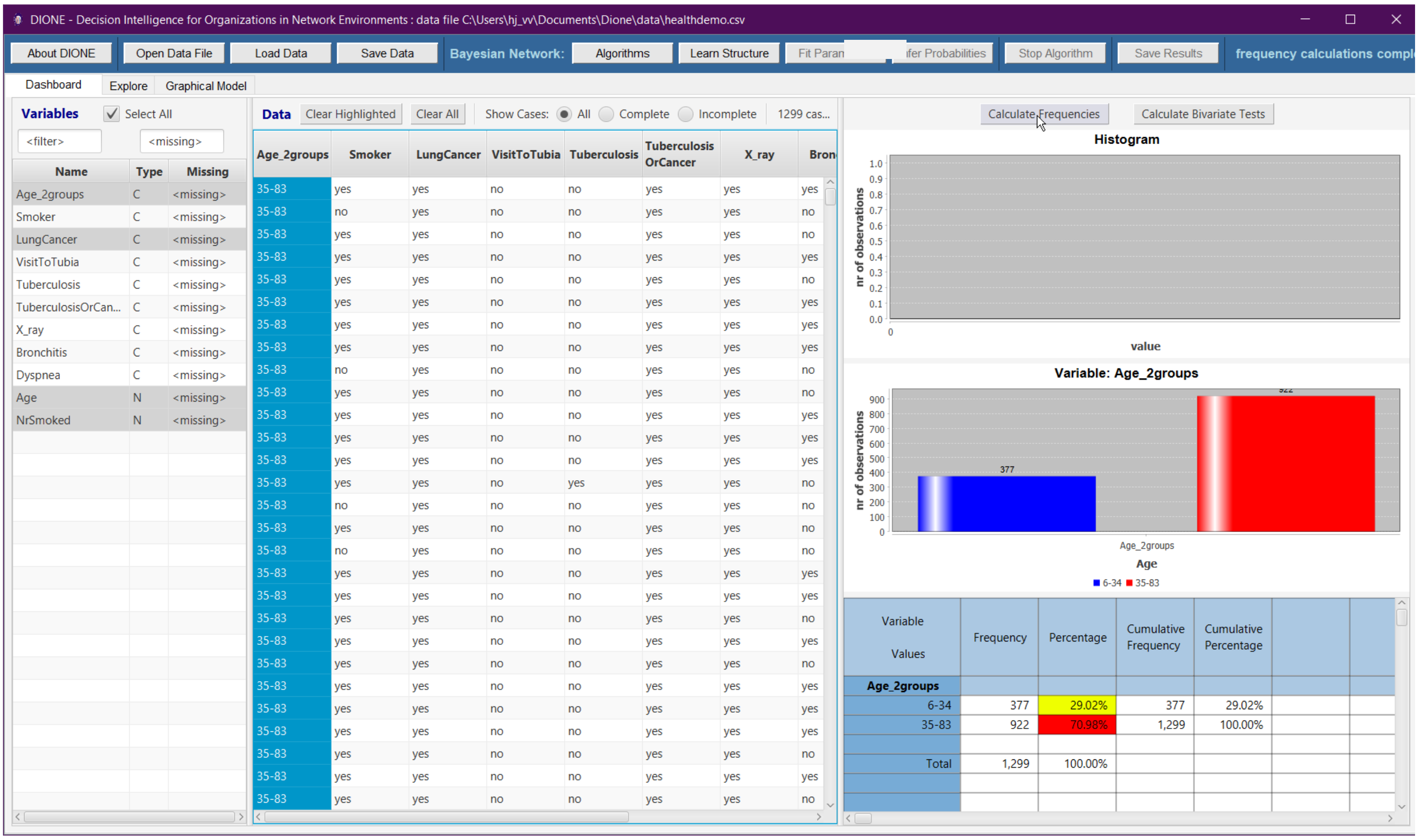
When you are satisfied with the new categories, perform the actual recoding by pressing button Recode Variable :



| Variable<br>Values | Freq |
|---|---|
| Age | |
| [6.0 - 13.7) | |
| [13.7 - 21.4) | |
| [21.4 - 29.1) | |
| [29.1 - 36.8) | |
| [36.8 - 44.5) | |
| [44.5 - 52.2) | |
| [52.2 - 59.9) | |
| [59.9 - 67.6) | |
| [67.6 - 75.3) | |
| [75.3 - 83.0] | |
| | |
| Total | |

This will create a new variable with the new name, with two categories (<35 and ≥35) and show it in the Dashboard tab, where the new variable appears at the top of the **Variables** table and in the first column of the **Data** table.

Press button Calculate Frequencies, press the `Enter` key or double-click in the new variable column to see its data:



**Variables** ☑ Select All

<filter> <missing>

| Name | Type | Missing |
|---|---|---|
| Age_2groups | C | <missing> |
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | N | <missing> |

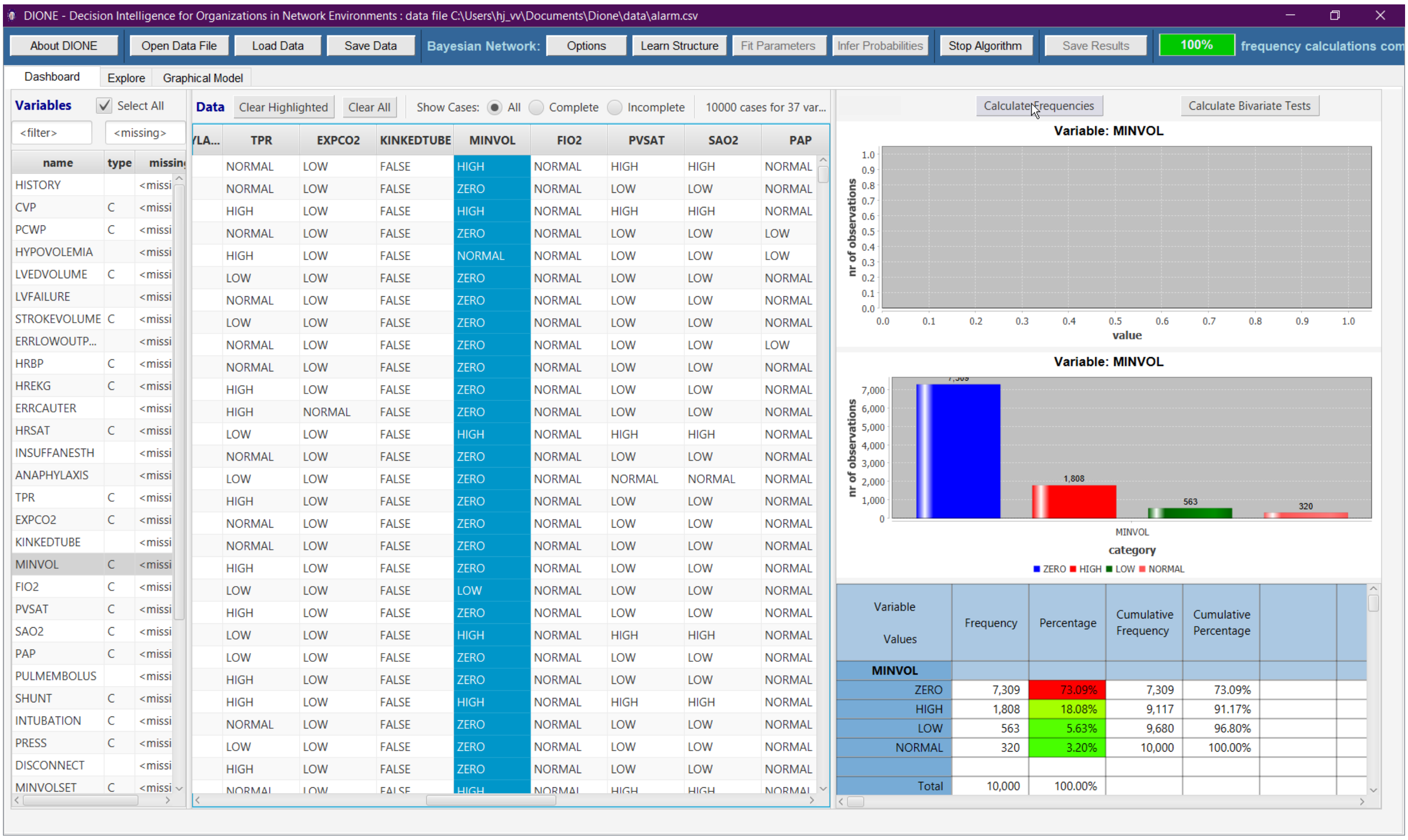
**Data** Clear Highlighted Clear All Show Cases: ● All ○ Complete ○ Incomplete 1299 cas...

| Age_2groups | Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bron |
|---|---|---|---|---|---|---|---|
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | no | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | no | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | yes | yes | yes | no |
| 35-83 | no | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | no |
| 35-83 | no | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no |

| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| **Age_2groups** | | | | |
| 6-34 | 377 | 29.02% | 377 | 29.02% |
| 35-83 | 922 | 70.98% | 1,299 | 100.00% |
| | | | | |
| Total | 1,299 | 100.00% | | |

## Categorical to Categorical

An example of recoding a categorical variable to a new categorical variable with different categories is variable **MINVOL** from data file `alarm.csv`. When **MINVOL** is selected in the Dashboard tab and its frequencies are displayed ( Calculate Frequencies , `Enter` key or double-click),

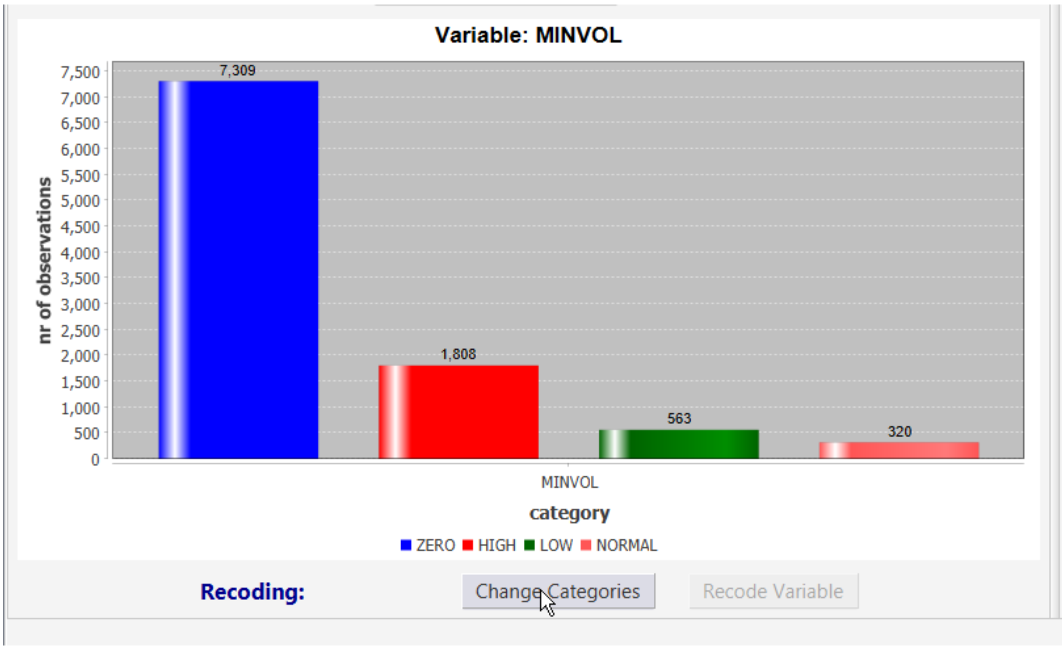


| name | type | missin |
|---|---|---|
| HISTORY | | <missi |
| CVP | C | <missi |
| PCWP | C | <missi |
| HYPOVOLEMIA | | <missi |
| LVEDVOLUME | C | <missi |
| LVFAILURE | | <missi |
| STROKEVOLUME | C | <missi |
| ERRLOWOUTP... | | <missi |
| HRBP | C | <missi |
| HREKG | C | <missi |
| ERRCAUTER | | <missi |
| HRSAT | C | <missi |
| INSUFFANESTH | | <missi |
| ANAPHYLAXIS | | <missi |
| TPR | C | <missi |
| EXPCO2 | C | <missi |
| KINKEDTUBE | | <missi |
| MINVOL | C | <missi |
| FIO2 | C | <missi |
| PVSAT | C | <missi |
| SAO2 | C | <missi |
| PAP | C | <missi |
| PULMEMBOLUS | | <missi |
| SHUNT | C | <missi |
| INTUBATION | C | <missi |
| PRESS | C | <missi |
| DISCONNECT | | <missi |
| MINVOLSET | C | <missi |

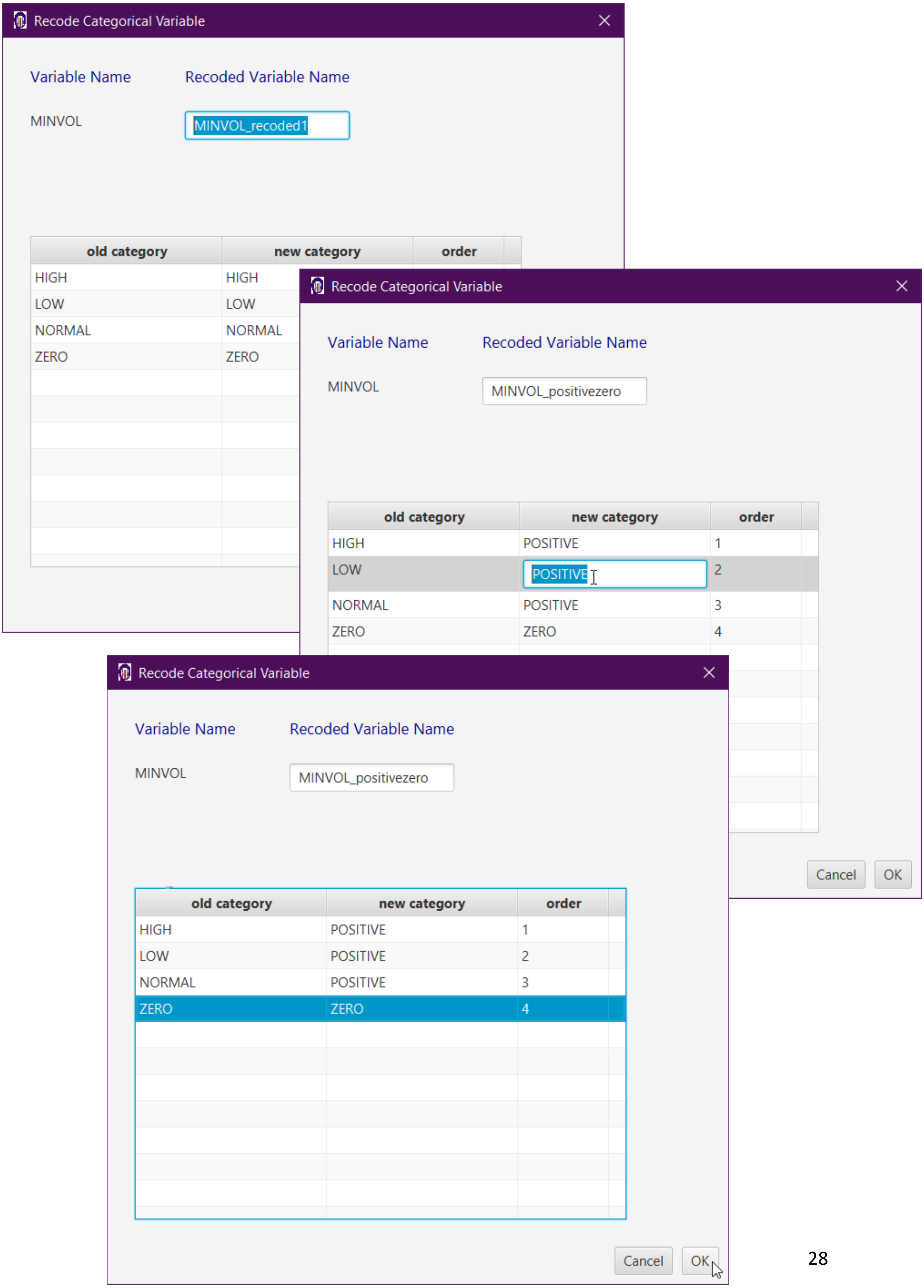
| /LA... | TPR | EXPCO2 | KINKEDTUBE | MINVOL | FIO2 | PVSAT | SAO2 | PAP |
|---|---|---|---|---|---|---|---|---|
| | NORMAL | LOW | FALSE | HIGH | NORMAL | HIGH | HIGH | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | HIGH | LOW | FALSE | HIGH | NORMAL | HIGH | HIGH | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | LOW |
| | HIGH | LOW | FALSE | NORMAL | NORMAL | LOW | LOW | LOW |
| | LOW | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | LOW | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | LOW |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | HIGH | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | HIGH | NORMAL | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | LOW | LOW | FALSE | HIGH | NORMAL | HIGH | HIGH | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | LOW | LOW | FALSE | ZERO | NORMAL | NORMAL | NORMAL | NORMAL |
| | HIGH | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | HIGH | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | LOW | LOW | FALSE | LOW | NORMAL | LOW | LOW | NORMAL |
| | HIGH | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | LOW | LOW | FALSE | HIGH | NORMAL | HIGH | HIGH | NORMAL |
| | LOW | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | HIGH | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | HIGH | LOW | FALSE | HIGH | NORMAL | HIGH | HIGH | NORMAL |
| | NORMAL | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | LOW | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | HIGH | LOW | FALSE | ZERO | NORMAL | LOW | LOW | NORMAL |
| | NORMAL | LOW | FALSE | HIGH | NORMAL | HIGH | HIGH | NORMAL |

| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| MINVOL | | | | |
| ZERO | 7,309 | 73.09% | 7,309 | 73.09% |
| HIGH | 1,808 | 18.08% | 9,117 | 91.17% |
| LOW | 563 | 5.63% | 9,680 | 96.80% |
| NORMAL | 320 | 3.20% | 10,000 | 100.00% |
| | | | | |
| Total | 10,000 | 100.00% | | |

the Explore tab shows this bar chart:



Now press button Change Categories to map the existing categories of the variable to the desired new categories and optionally give a new name for the recoded variable:



Recode Categorical Variable

Variable Name: MINVOL

Recoded Variable Name: MINVOL_recoded1

| old category | new category | order |
|---|---|---|
| HIGH | HIGH | |
| LOW | LOW | |
| NORMAL | NORMAL | |
| ZERO | ZERO | |

Recode Categorical Variable

Variable Name: MINVOL

Recoded Variable Name: MINVOL_positivezero

| old category | new category | order |
|---|---|---|
| HIGH | POSITIVE | 1 |
| LOW | POSITIVE | 2 |
| NORMAL | POSITIVE | 3 |
| ZERO | ZERO | 4 |

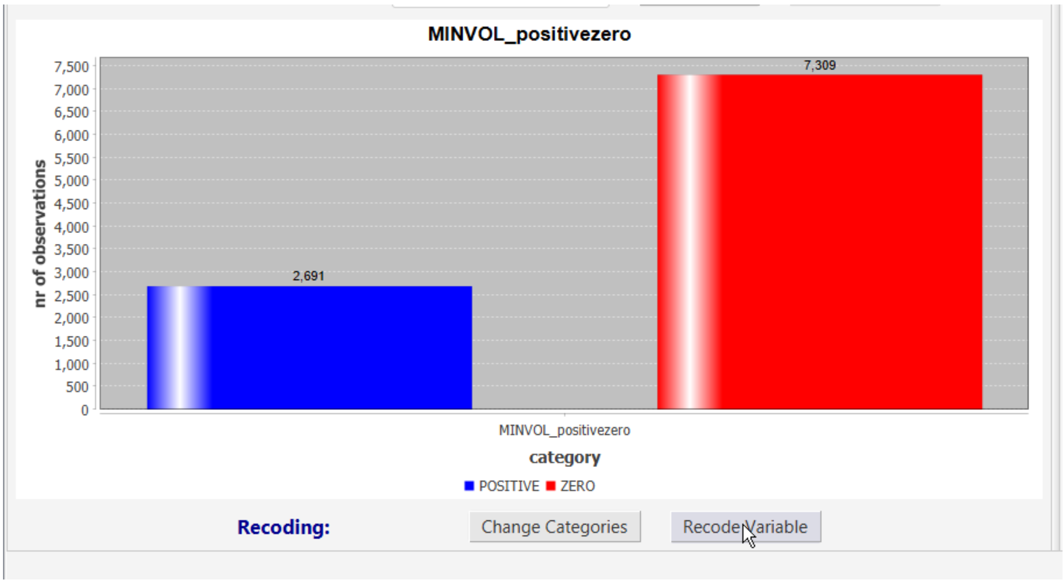
Recode Categorical Variable

Variable Name: MINVOL

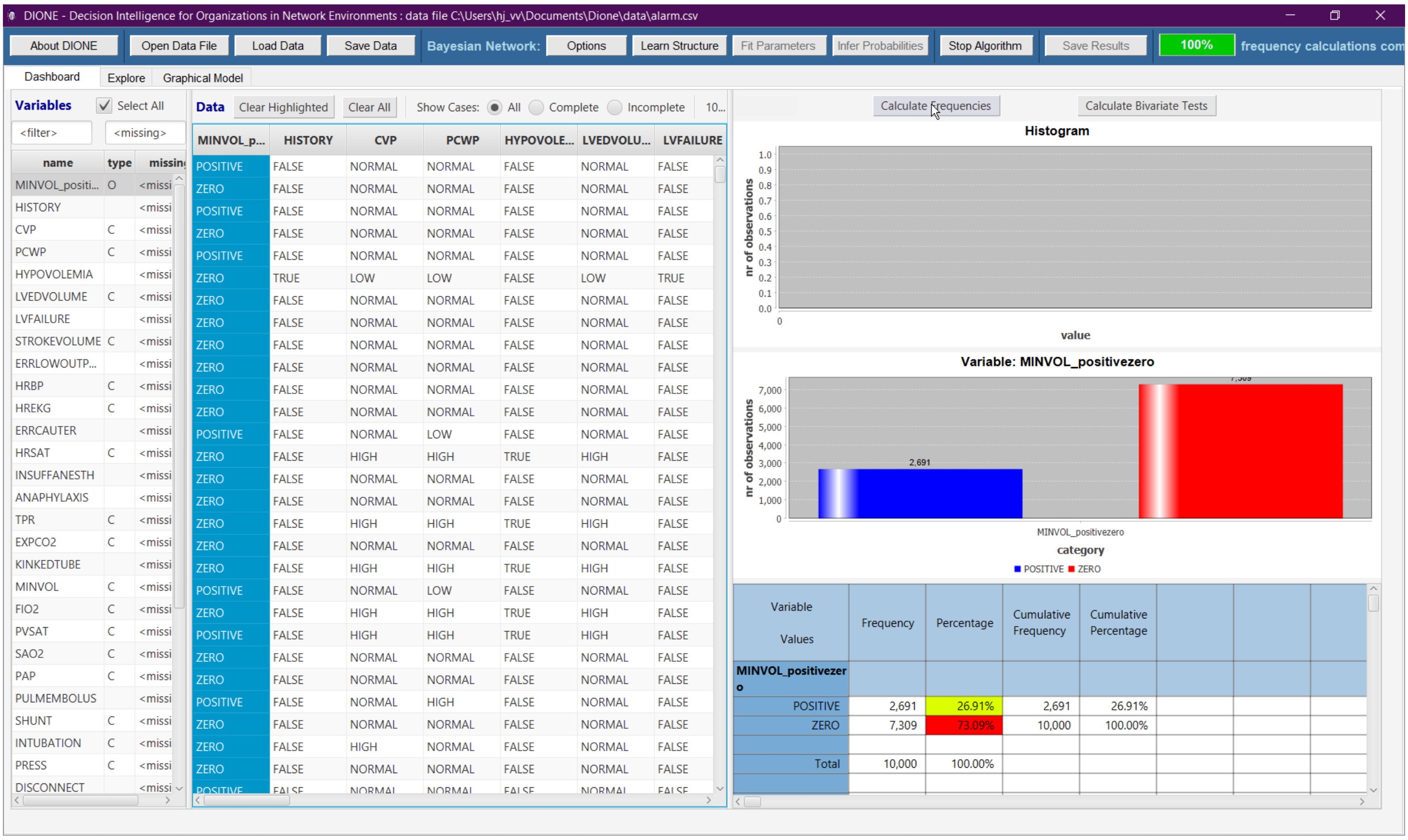
Recoded Variable Name: MINVOL_positivezero

| old category | new category | order |
|---|---|---|
| HIGH | POSITIVE | 1 |
| LOW | POSITIVE | 2 |
| NORMAL | POSITIVE | 3 |
| ZERO | ZERO | 4 |

When you are satisfied with the proposed recoding, press button Recode Variable next to the Change Categories button to actually create the new recoded variable:



This will create a new variable with the new name, with two categories (`POSITIVE` and `ZERO`), and show it in the Dashboard tab, where it appears at the top of the **Variables** table and in the first column of the **Data** table. Press button Calculate Frequencies , the `Enter` key or double-click in the new variable column to see its data:



| name | type | missin |
|---|---|---|
| MINVOL_positi... | O | <missi |
| HISTORY | | <missi |
| CVP | C | <missi |
| PCWP | C | <missi |
| HYPOVOLEMIA | | <missi |
| LVEDVOLUME | C | <missi |
| LVFAILURE | | <missi |
| STROKEVOLUME | C | <missi |
| ERRLOWOUTP... | | <missi |
| HRBP | C | <missi |
| HREKG | C | <missi |
| ERRCAUTER | | <missi |
| HRSAT | C | <missi |
| INSUFFANESTH | | <missi |
| ANAPHYLAXIS | | <missi |
| TPR | C | <missi |
| EXPCO2 | C | <missi |
| KINKEDTUBE | | <missi |
| MINVOL | C | <missi |
| FIO2 | C | <missi |
| PVSAT | C | <missi |
| SAO2 | C | <missi |
| PAP | C | <missi |
| PULMEMBOLUS | | <missi |
| SHUNT | C | <missi |
| INTUBATION | C | <missi |
| PRESS | C | <missi |
| DISCONNECT | | <missi |

| MINVOL_p... | HISTORY | CVP | PCWP | HYPOVOLE... | LVEDVOLU... | LVFAILURE |
|---|---|---|---|---|---|---|
| POSITIVE | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| POSITIVE | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| POSITIVE | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | TRUE | LOW | LOW | FALSE | LOW | TRUE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| POSITIVE | FALSE | NORMAL | LOW | FALSE | NORMAL | FALSE |
| ZERO | FALSE | HIGH | HIGH | TRUE | HIGH | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | HIGH | HIGH | TRUE | HIGH | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | HIGH | HIGH | TRUE | HIGH | FALSE |
| POSITIVE | FALSE | NORMAL | LOW | FALSE | NORMAL | FALSE |
| ZERO | FALSE | HIGH | HIGH | TRUE | HIGH | FALSE |
| POSITIVE | FALSE | HIGH | HIGH | TRUE | HIGH | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| POSITIVE | FALSE | NORMAL | HIGH | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | HIGH | NORMAL | FALSE | NORMAL | FALSE |
| ZERO | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |
| POSITIVE | FALSE | NORMAL | NORMAL | FALSE | NORMAL | FALSE |

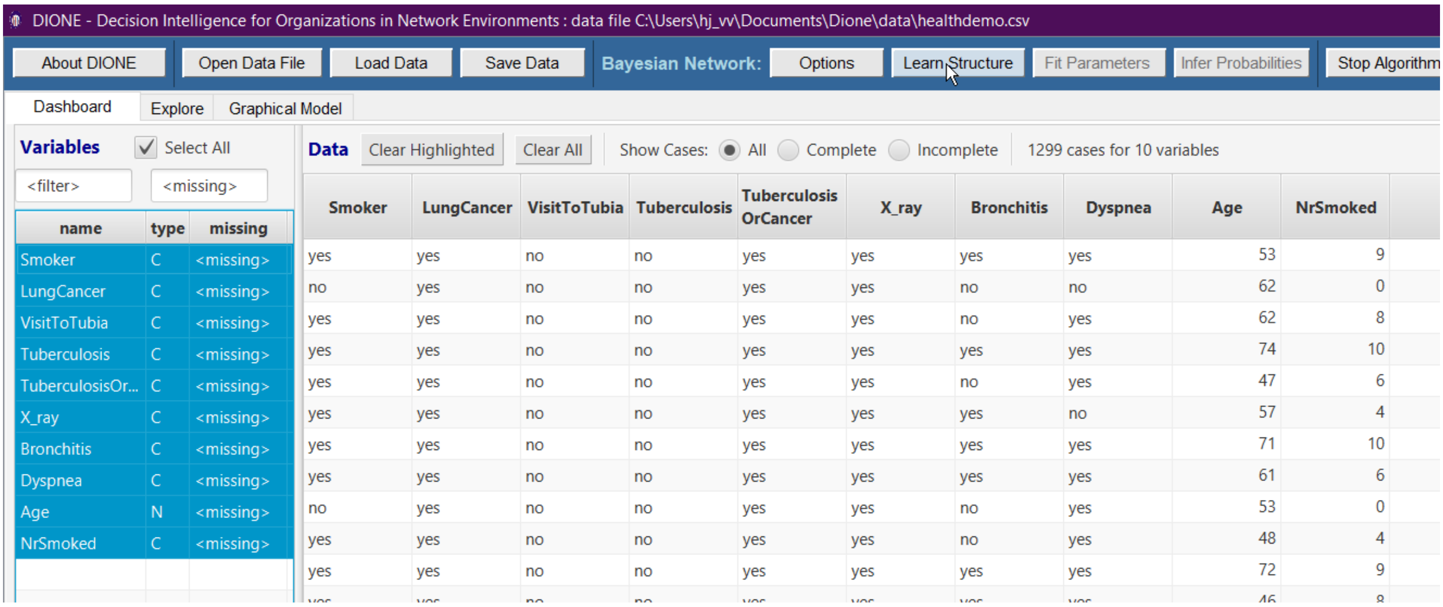
| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage | | | |
|---|---|---|---|---|---|---|---|
| MINVOL_positivezero | | | | | | | |
| POSITIVE | 2,691 | 26.91% | 2,691 | 26.91% | | | |
| ZERO | 7,309 | 73.09% | 10,000 | 100.00% | | | |
| | | | | | | | |
| Total | 10,000 | 100.00% | | | | | |
| | | | | | | | |

# Learn Bayesian Networks

In the Dashboard tab, ensure that the variables you want to analyse are selected in the **Variables** table and, by pressing the Load Data button or the `Enter` key, loaded and shown in the **Data** table. If you want to change the variables to analyse, change the selected variables as explained under **Load Data**.

## Learn Network Structure

Now press button Learn Structure to learn a Bayesian network structure from the selected data.



☑ Select All

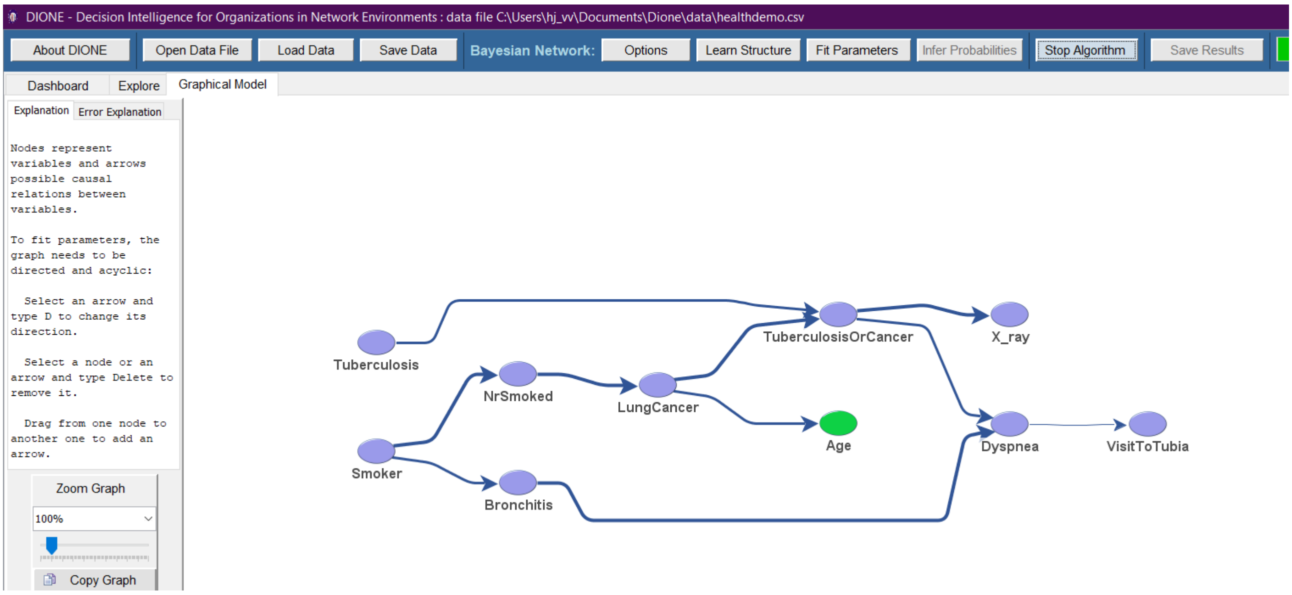
<filter> <missing>

| name | type | missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOr... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | C | <missing> |

Show Cases: ● All ○ Complete ○ Incomplete 1299 cases for 10 variables

| Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis | Dyspnea | Age | NrSmoked |
|---|---|---|---|---|---|---|---|---|---|
| yes | yes | no | no | yes | yes | yes | yes | 53 | 9 |
| no | yes | no | no | yes | yes | no | no | 62 | 0 |
| yes | yes | no | no | yes | yes | no | yes | 62 | 8 |
| yes | yes | no | no | yes | yes | yes | yes | 74 | 10 |
| yes | yes | no | no | yes | yes | no | yes | 47 | 6 |
| yes | yes | no | no | yes | yes | yes | no | 57 | 4 |
| yes | yes | no | no | yes | yes | yes | yes | 71 | 10 |
| yes | yes | no | no | yes | yes | yes | yes | 61 | 6 |
| no | yes | no | no | yes | yes | no | yes | 53 | 0 |
| yes | yes | no | no | yes | yes | no | yes | 48 | 4 |
| yes | yes | no | no | yes | yes | yes | yes | 72 | 9 |
| yes | yes | no | no | yes | yes | yes | yes | 46 | 8 |

The result should look something like this in the Graphical Model tab:

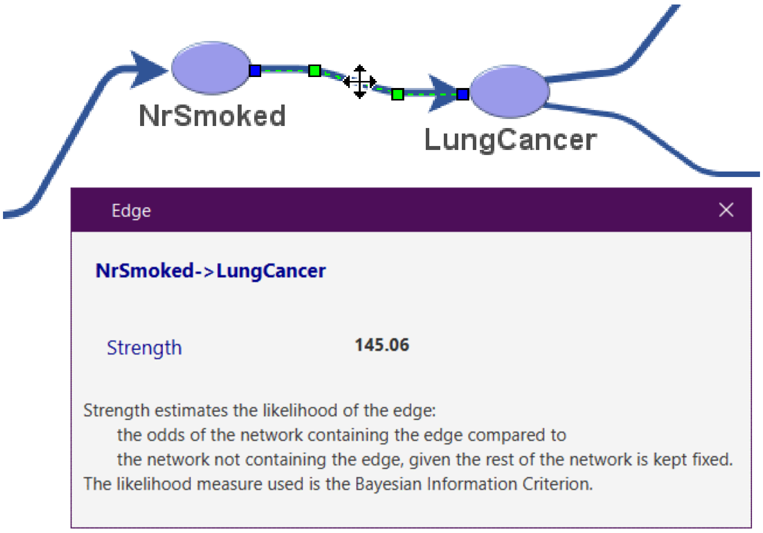


```
Nodes represent
variables and arrows
possible causal
relations between
variables.

To fit parameters, the
graph needs to be
directed and acyclic:

  Select an arrow and
type D to change its
direction.

  Select a node or an
arrow and type Delete to
remove it.

  Drag from one node to
another one to add an
arrow.
```

In the learned Bayesian network, nodes represent variables and arrows represent possible causal relations between variables. Directed lines, with arrows, represent possible causal relations including the direction of causality, while any undirected lines, without arrows, represent possible causal relations with unspecified causal direction. If there is no significant association between two variables, there is no line between their nodes. The default structure learning algorithm, tabu search, produces a directed graph with only arrows.

The algorithm also calculates edge strengths. The strength of an edge estimates the likelihood of the edge: the odds of the network containing the edge compared to the network not containing the edge, given the rest of the network is kept fixed. The likelihood measure used is the Bayesian Information Criterion.

Edge strengths are indicated by the thickness and colour of the arrows. The thicker an arrow, the stronger is the evidence for the causal connection. Edge strengths can be inspected by right-clicking on an arrow:



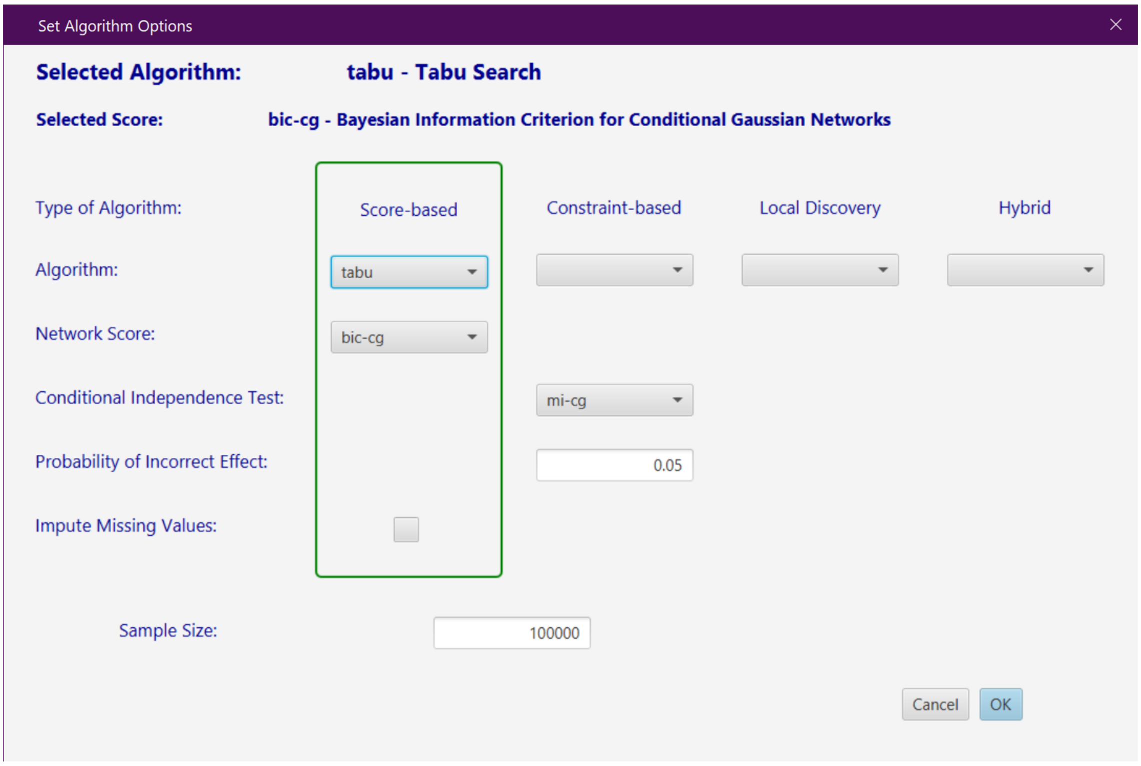
NrSmoked->LungCancer

Strength 145.06

Strength estimates the likelihood of the edge:
the odds of the network containing the edge compared to
the network not containing the edge, given the rest of the network is kept fixed.
The likelihood measure used is the Bayesian Information Criterion.

In this example the strength of the connection between **NrSmoked** and **LungCancer** is over 145, indication that the network with this connection is much more likely than a network without the connection and otherwise the same.

The default structure learning algorithm is tabu search, but other structure learning algorithms can be used by selecting one in the Set Algorithm Options dialog accessed by pressing the Algorithms button:



Set Algorithm Options

**Selected Algorithm:** **tabu - Tabu Search**

**Selected Score:** **bic-cg - Bayesian Information Criterion for Conditional Gaussian Networks**

| Type of Algorithm: | Score-based | Constraint-based | Local Discovery | Hybrid |
|---|---|---|---|---|
| Algorithm: | tabu | | | |
| Network Score: | bic-cg | | | |
| Conditional Independence Test: | | mi-cg | | |
| Probability of Incorrect Effect: | | 0.05 | | |
| Impute Missing Values: | □ | | | |

Sample Size: 100000

DIONE uses the R package bnlearn for structure learning and parameter fitting. Supported structure learning algorithms are described in the bnlearn manual on pages 100 – 102 at bookmark structure-learning. Structure learning algorithms can be Score-based, Constraint-based, Local Discovery or Hybrid algorithms. To use a specific algorithm, select it in the Algorithm drop down list under the relevant category.

Score-based algorithms use a network score to compare solutions. Possible network scores are described in the bnlearn manual on pages 81 – 82 at bookmark network-scores. To use a specific score, select it in the Network Score drop down list.

Constraint-based algorithms and Local Discovery algorithms use a conditional independence test to find solutions. Possible tests are described in the bnlearn manual on pages 66 – 67 at bookmark independence-tests. To use a specific test, select it in the Conditional Independence Test drop down list.

Hybrid algorithms use a network score as well as a conditional independence test. When a Hybrid algorithm is selected in the Algorithm drop down list under Hybrid, the algorithm uses the Network Score selected under Score-based and the Conditional Independence Test selected under Constraint-based. When the rsmax2 Hybrid algorithm is selected, it combines the algorithms selected under Score-based and under Constraint-based.

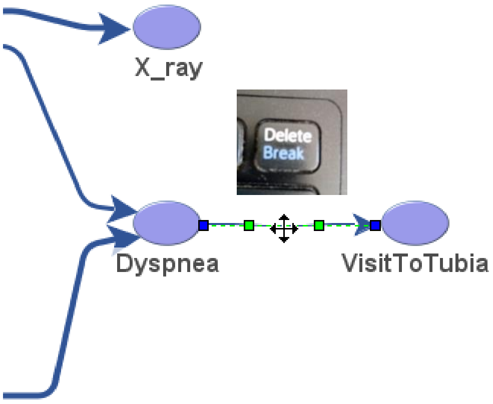
## Edit Network Structure

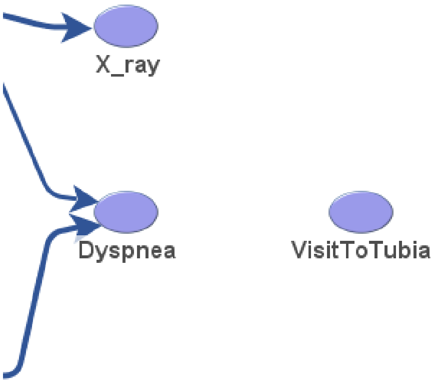
The next step in the analysis will be the parameter fitting algorithm, which calculates detailed information for all nodes and arrows, such as conditional probabilities and/or regression coefficients quantifying the causal relationships between variables. Before continuing with parameter fitting, you may wish to edit the network structure to reflect knowledge or intuition you already have about causal connections between variables. You can move or delete nodes, add or delete arrows or reverse their direction.

Adding an arrow or changing direction of an arrow can possibly result in a red arrow, meaning that a network with that connection is less likely than a network without it.

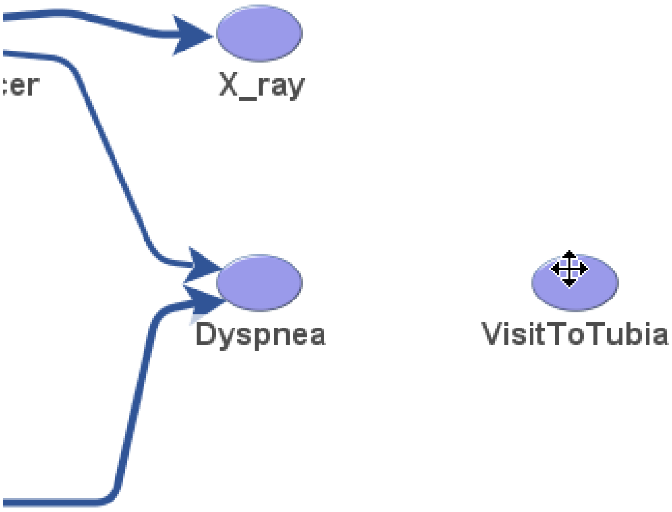
For example, in the learned network shown on page 20 the arrow between variables **Dyspnea** and **VisitToTubia** does not seem to reflect a likely causal connection. Let us say we want to remove it and instead postulate a causal connection between **VisitToTubia** and **Tuberculosis**.

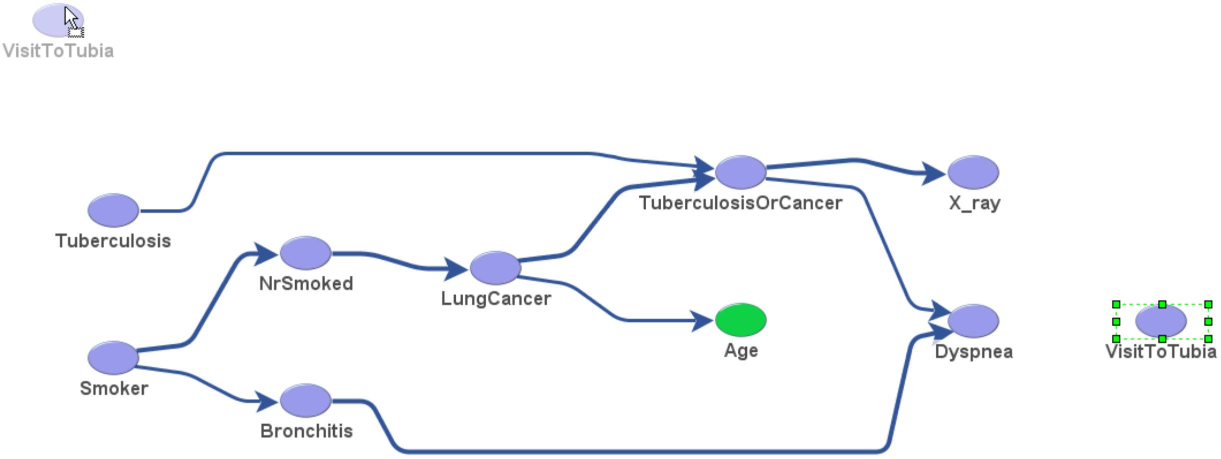
First, we delete the arrow between **Dyspnea** and **VisitToTubia** with the `Delete` key:

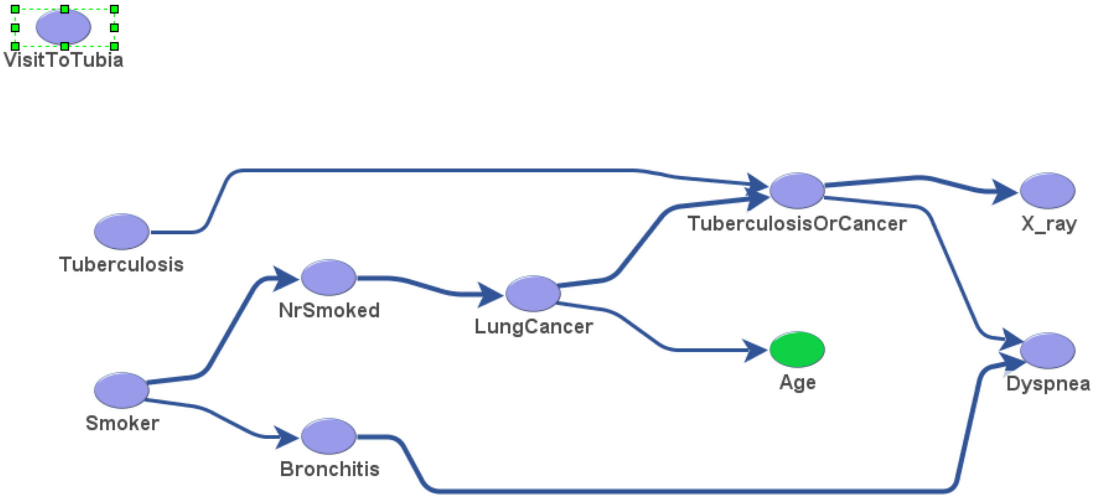




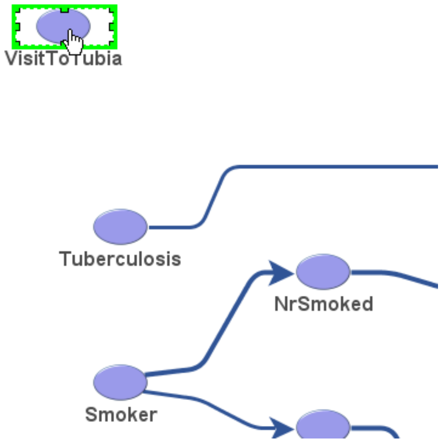
Now, to keep the diagram tidy, we move the node of variable **VisitToTubia** to a position near the **Tuberculosis** node. To move a node, hover over it with the mouse to show a black cross and drag it to the desired location:

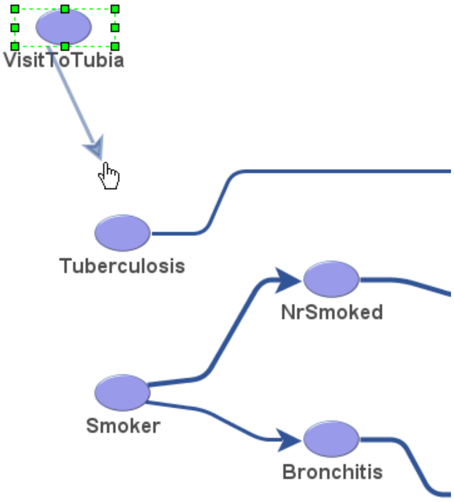

:er
X_ray
Dyspnea
VisitToTubia

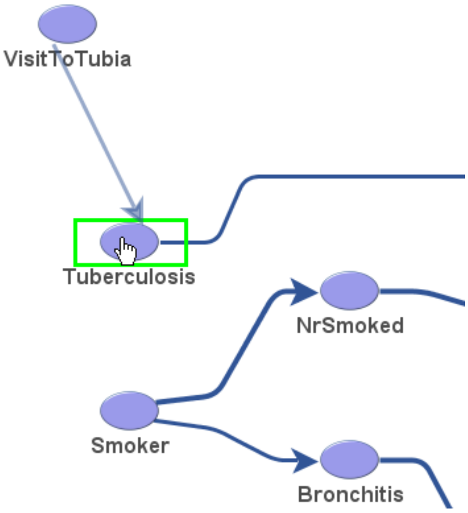

VisitToTubia
Tuberculosis
TuberculosisOrCancer
X_ray
NrSmoked
LungCancer
Age
Dyspnea
VisitToTubia
Smoker
Bronchitis

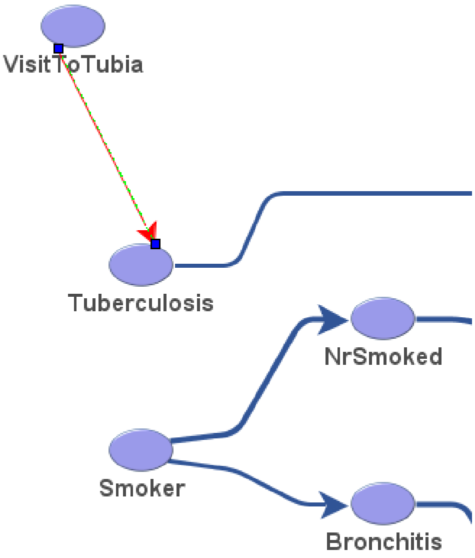

VisitToTubia
Tuberculosis
TuberculosisOrCancer
X_ray
NrSmoked
LungCancer
Age
Dyspnea
Smoker
Bronchitis

To add a new arrow between the two nodes **VisitToTubia** and **Tuberculosis**, first hover with the cursor over the node that is considered the cause (**VisitToTubia**) and where the arrow is to start, such that the cursor is a hand and there is a green frame around the node. Now press the left mouse button and, while holding it, move the mouse to draw an arrow to the node that is considered the effect (**Tuberculosis**) and where the arrow is to end. When there is a green frame around the effect node, release the mouse button. The new arrow has now been added:



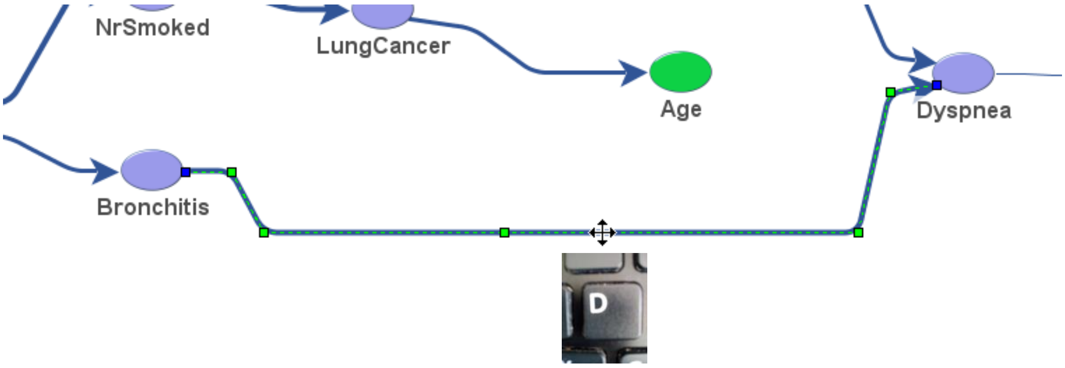


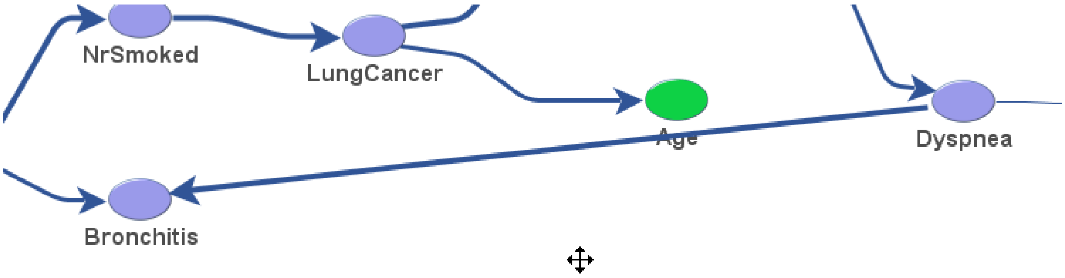




Note that the arrow between nodes **VisitToTubia** and **Tuberculosis** , which has been manually added, is red, meaning that the causal relationship between variables **VisitToTubia** and **Tuberculosis** is not likely, given the data.

You can also reverse the direction of directed arrows if their direction does not make sense to you, or to eliminate cycles. To reverse the direction of an arrow, select it by clicking on the line of the arrow and push the **D** key on your keyboard:





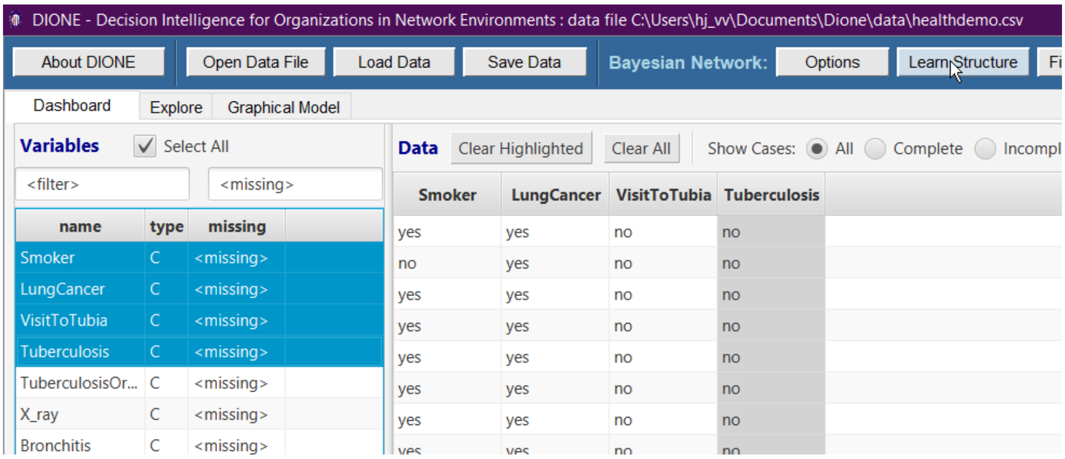
If undirected arrows are present in the learned network, you can give them a direction in the same way, pushing the **D** key to give an arrow a direction and possibly pushing it again to reverse the direction.

## Fit Parameters

More information about causal relations between variables in the network can be gained by fitting parameters of the network to obtain conditional probability tables and regression coefficients, showing the (probabilities of) values of a node given values of its parent nodes. Parameters are estimated with a Maximum Likelihood algorithm.

When you are done editing network connections and the graph is completely directed, i.e. there are no undirected edges present, and the graph does not have cycles, you can calculate conditional probabilities, marginal probabilities and regression coefficients for numerical variables by pressing button Fit Parameters .

For clarity in the example shown here structure learning has been done with only four variables:

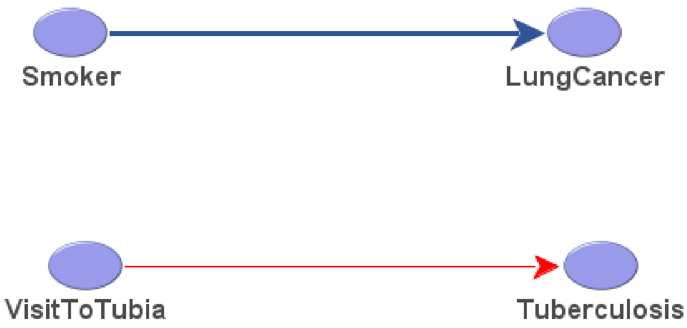


DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemo.csv

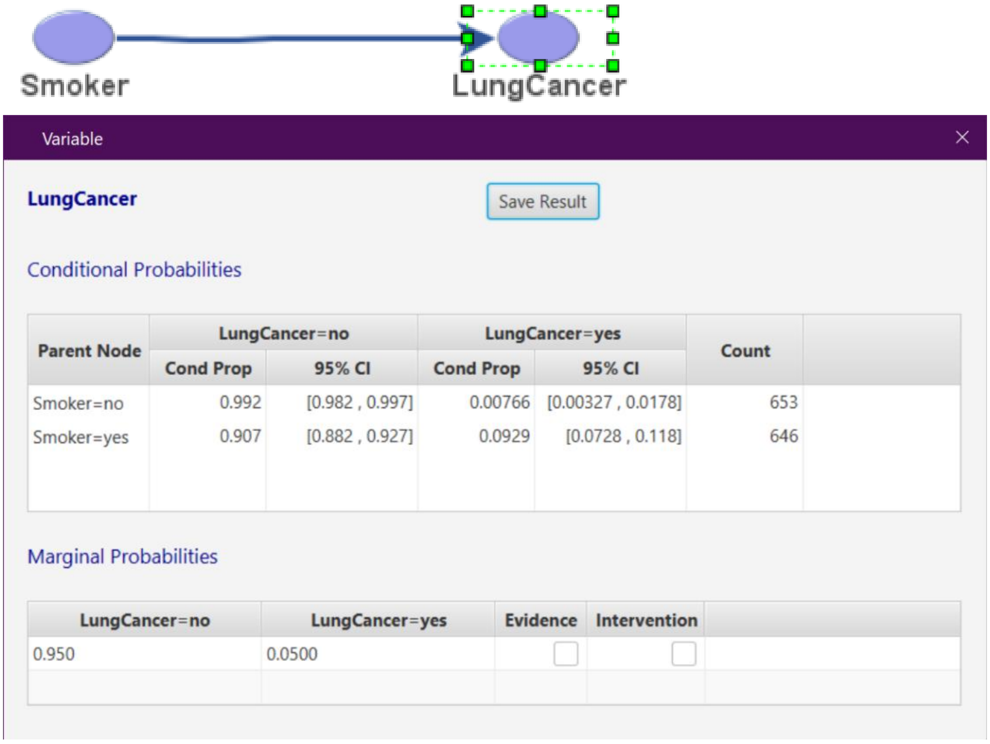
| name | type | missing |
|---|---|---|
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOr... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |

| Smoker | LungCancer | VisitToTubia | Tuberculosis |
|---|---|---|---|
| yes | yes | no | no |
| no | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |
| yes | yes | no | no |

After some editing, we have this simple network:



When the parameter fitting algorithm has finished, right-click on any node of a categorical variable (shown in blue) to see its conditional probability table, showing conditional probabilities given the values of its parents, along with 95 % confidence intervals for these conditional probabilities. At the bottom of the dialog marginal probabilities of the node are shown:



LungCancer

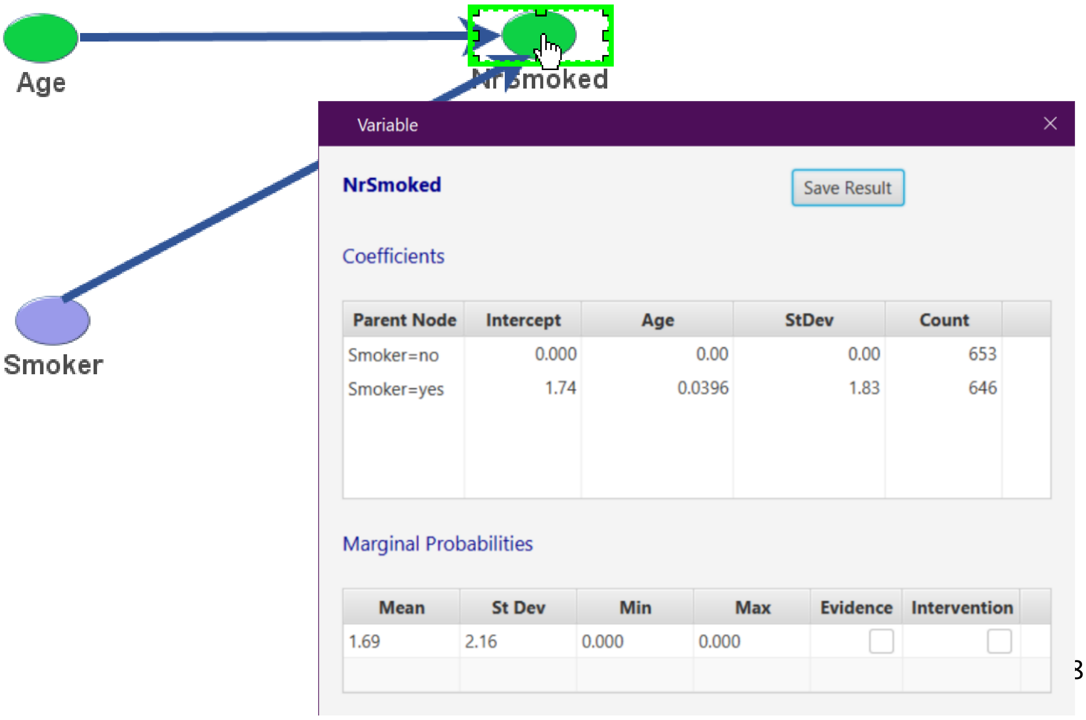
Conditional Probabilities

| Parent Node | LungCancer=no | | LungCancer=yes | | Count |
|---|---|---|---|---|---|
| | Cond Prop | 95% CI | Cond Prop | 95% CI | |
| Smoker=no | 0.992 | [0.982 , 0.997] | 0.00766 | [0.00327 , 0.0178] | 653 |
| Smoker=yes | 0.907 | [0.882 , 0.927] | 0.0929 | [0.0728 , 0.118] | 646 |

Marginal Probabilities

| LungCancer=no | LungCancer=yes | Evidence | Intervention |
|---|---|---|---|
| 0.950 | 0.0500 | ☐ | ☐ |

For a node of a numerical variable (shown in green), coefficients of regression equations are displayed, along with the standard deviation of the residuals. In this example, variable **NrSmoked** depends on the two variables **Age** and **Smoker**. **Smoker** has two possible values, so there are regression equations reflecting the dependence of **NrSmoked** on **Age** for each of the values of **Smoker**:



NrSmoked

Coefficients

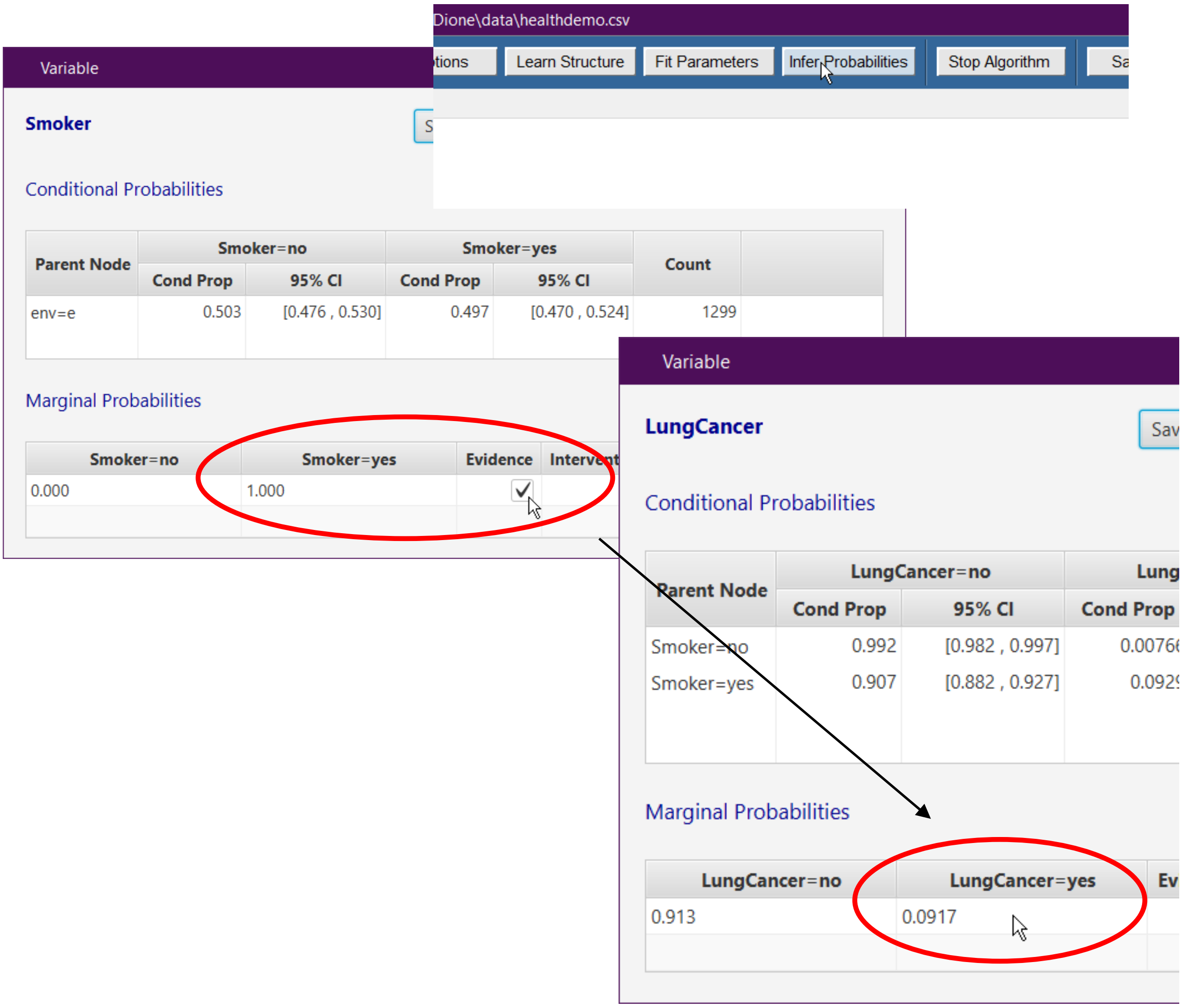
| Parent Node | Intercept | Age | StDev | Count |
|---|---|---|---|---|
| Smoker=no | 0.000 | 0.00 | 0.00 | 653 |
| Smoker=yes | 1.74 | 0.0396 | 1.83 | 646 |

Marginal Probabilities

| Mean | St Dev | Min | Max | Evidence | Intervention |
|---|---|---|---|---|---|
| 1.69 | 2.16 | 0.000 | 0.000 | ☐ | ☐ |

## Infer Probabilities

Now you can perform inference on the Bayesian network by setting evidence on one or more nodes, then pressing button Infer Probabilities to recompute marginal probabilities and see the effect of the evidence on the other nodes:



**Smoker**

Conditional Probabilities

| Parent Node | Smoker=no | | Smoker=yes | | Count |
|---|---|---|---|---|---|
| | Cond Prop | 95% CI | Cond Prop | 95% CI | |
| env=e | 0.503 | [0.476 , 0.530] | 0.497 | [0.470 , 0.524] | 1299 |

Marginal Probabilities

| Smoker=no | Smoker=yes | Evidence | Intervent |
|---|---|---|---|
| 0.000 | 1.000 | ☑ | |

**LungCancer**

Conditional Probabilities

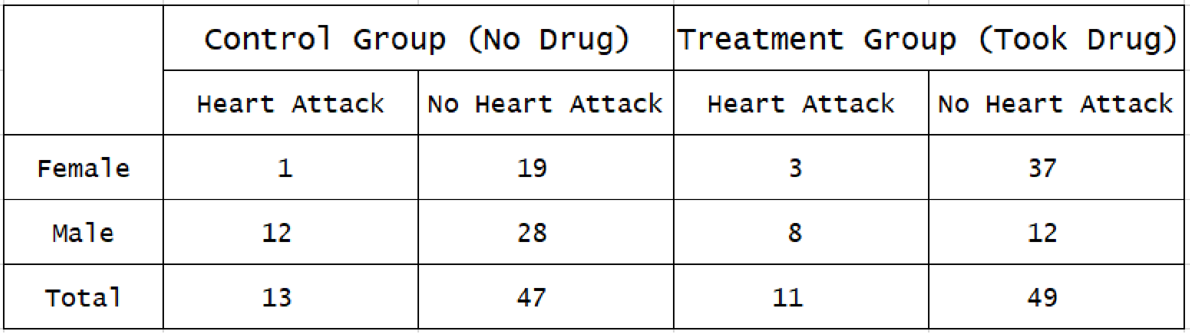
| Parent Node | LungCancer=no | | Lung |
|---|---|---|---|
| | Cond Prop | 95% CI | Cond Prop |
| Smoker=no | 0.992 | [0.982 , 0.997] | 0.0076 |
| Smoker=yes | 0.907 | [0.882 , 0.927] | 0.092 |

Marginal Probabilities

| LungCancer=no | LungCancer=yes | Ev |
|---|---|---|
| 0.913 | 0.0917 | |

In this example the probability of lung cancer has increased from 5 % to over 9 % on the evidence that someone is a smoker.

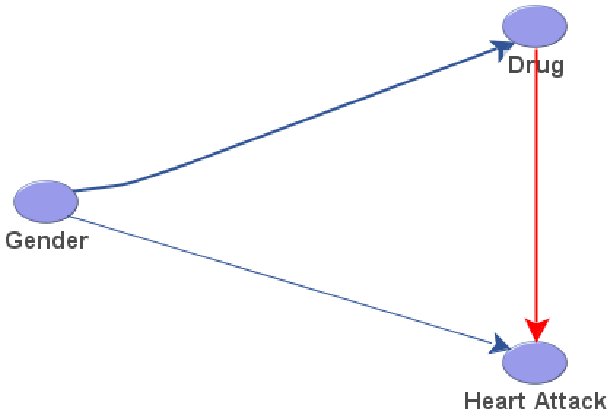
We can also infer the effects of interventions, as illustrated by the following example related to the famous Simpson paradox. The Simpson paradox occurs in these fictitious data on gender, taking of a drug and risk of a heart attack (from Pearl & Mackenzie 2018):



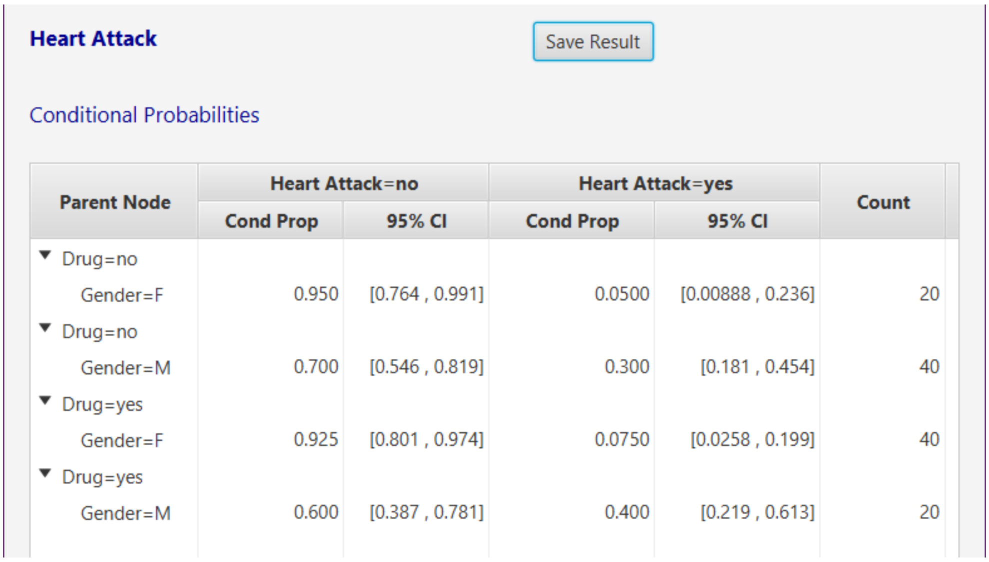
| | Control Group (No Drug) | | Treatment Group (Took Drug) | |
|---|---|---|---|---|
| | Heart Attack | No Heart Attack | Heart Attack | No Heart Attack |
| Female | 1 | 19 | 3 | 37 |
| Male | 12 | 28 | 8 | 12 |
| Total | 13 | 47 | 11 | 49 |

From these data it seems that the drug increases the risk of a heart attack for women, as 3/40 (7.5%) > 1/20 (5%), it also increases the risk for men, as 8/20 (40%) > 12/40 (30%), yet it decreases the risk for the population as a whole, as 11/60 (18.33%) < 13/60 (21.67%).

This is shown in **DIONE** by the data in file simpson.csv: Learn Structure and add an arrow to the learned network



After Fit Parameters right-click on the **Heart Attack** node to see its conditional probabilities:

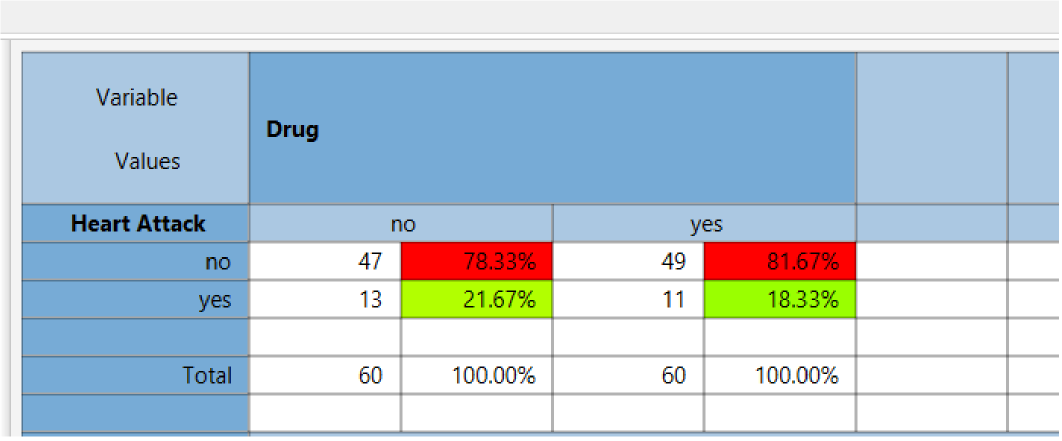


**Heart Attack**

Conditional Probabilities

| Parent Node | Heart Attack=no | | Heart Attack=yes | | Count |
|---|---|---|---|---|---|
| | Cond Prop | 95% CI | Cond Prop | 95% CI | |
| Drug=no<br>Gender=F | 0.950 | [0.764 , 0.991] | 0.0500 | [0.00888 , 0.236] | 20 |
| Drug=no<br>Gender=M | 0.700 | [0.546 , 0.819] | 0.300 | [0.181 , 0.454] | 40 |
| Drug=yes<br>Gender=F | 0.925 | [0.801 , 0.974] | 0.0750 | [0.0258 , 0.199] | 40 |
| Drug=yes<br>Gender=M | 0.600 | [0.387 , 0.781] | 0.400 | [0.219 , 0.613] | 20 |

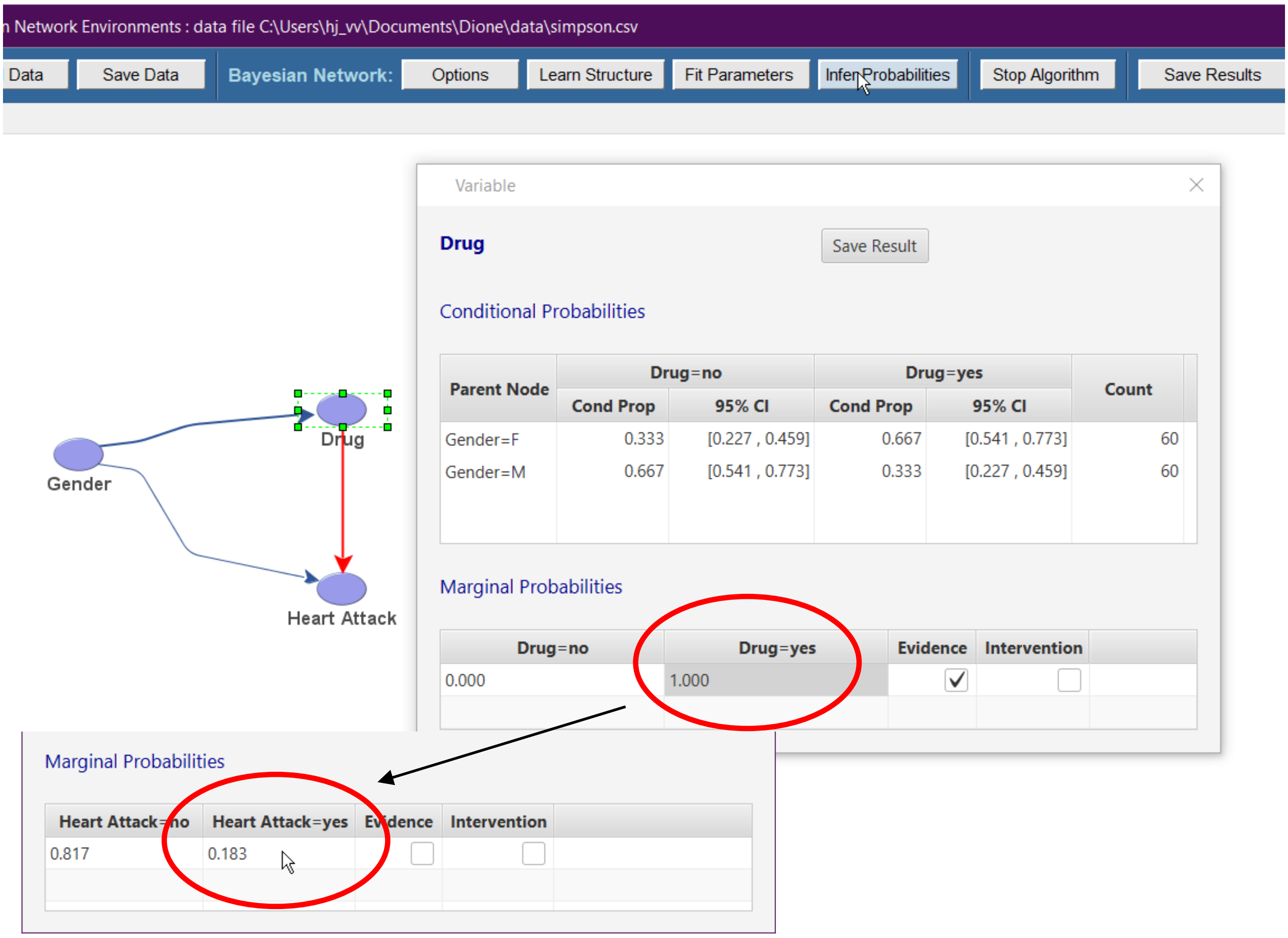
The conditional probabilities of getting a heart attack according to gender are the same as the percentages in the table. Yet, when comparing variables Drug and Heart Attack in the Dashboard tab, we see these percentages for the population:



| Variable<br>Values | Drug | | | |
|---|---|---|---|---|
| **Heart Attack** | no | | yes | |
| no | 47 | 78.33% | 49 | 81.67% |
| yes | 13 | 21.67% | 11 | 18.33% |
| | | | | |
| Total | 60 | 100.00% | 60 | 100.00% |

The paradox is resolved by realizing that **Gender** is a confounding variable that influences both the risk of a heart attack and the probability of taking the drug: women have less risk of a heart attack, but also take the drug more often than men. To correctly estimate the effect of the drug on the risk of a heart attack, we must adjust for gender and estimate the risk of a heart attack when taking the drug by averaging the percentages for men and women.

The paradox is illustrated in **DIONE** by inferring the probability of a heart attack given observational *evidence* of taking the drug, compared to the effect of an *intervention* to take the drug. Setting evidence that the drug has been taken and inferring probabilities gives the following results, with a probability of a heart attack of 18.3% when taking the drug and 21.7% when not taking the drug:



Drug

Conditional Probabilities

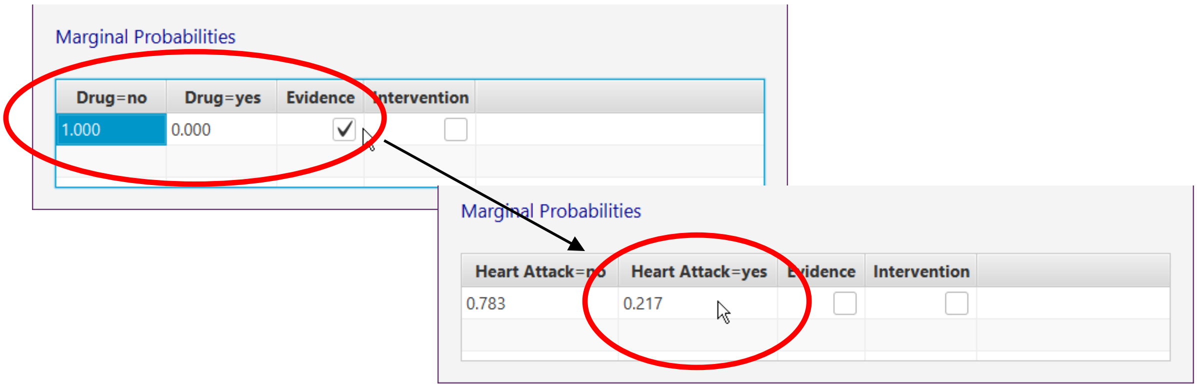
| Parent Node | Drug=no |  | Drug=yes |  | Count |
|---|---|---|---|---|---|
|  | Cond Prop | 95% CI | Cond Prop | 95% CI |  |
| Gender=F | 0.333 | [0.227 , 0.459] | 0.667 | [0.541 , 0.773] | 60 |
| Gender=M | 0.667 | [0.541 , 0.773] | 0.333 | [0.227 , 0.459] | 60 |

Marginal Probabilities

| Drug=no | Drug=yes | Evidence | Intervention |
|---|---|---|---|
| 0.000 | 1.000 | ☑ | □ |

Marginal Probabilities

| Heart Attack=no | Heart Attack=yes | Evidence | Intervention |
|---|---|---|---|
| 0.817 | 0.183 | □ | □ |



Marginal Probabilities

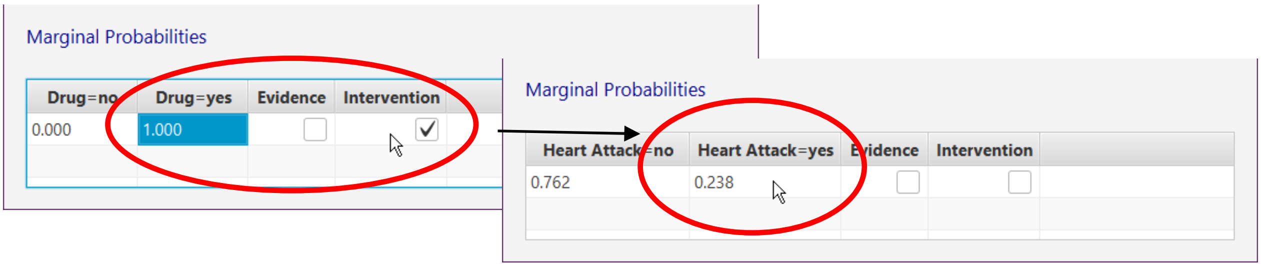
| Drug=no | Drug=yes | Evidence | Intervention |
|---|---|---|---|
| 1.000 | 0.000 | ☑ | □ |

Marginal Probabilities

| Heart Attack=no | Heart Attack=yes | Evidence | Intervention |
|---|---|---|---|
| 0.783 | 0.217 | □ | □ |

The evidence means we have only observed that the drug has been taken without doing anything, and the inference will include the effect of the confounding variable **Gender**, so the incorrect conclusion could be drawn that the drug is beneficial for the population.

When setting an *intervention* to take the drug, evidence is also set that the drug has been taken, but now the effect of confounding is eliminated by disregarding the arrow from **Gender** to **Drug** in the network. We are assuming the drug is now given to people, without considering gender. This gives the following inference results:

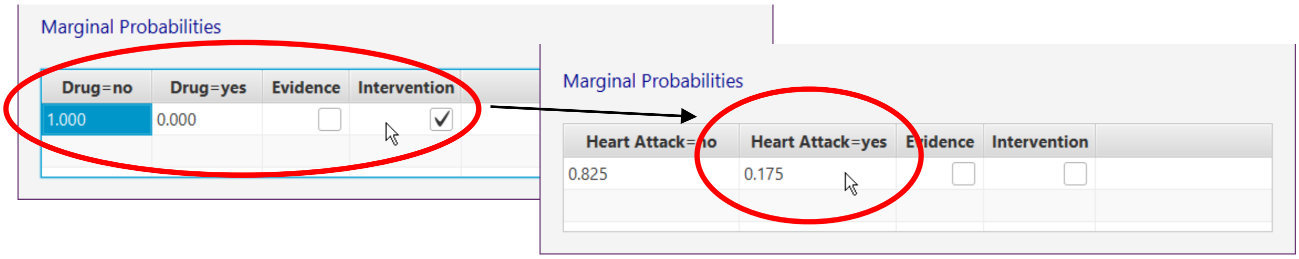


Marginal Probabilities

| Drug=no | Drug=yes | Evidence | Intervention |
|---|---|---|---|
| 0.000 | 1.000 | □ | ☑ |

Marginal Probabilities

| Heart Attack=no | Heart Attack=yes | Evidence | Intervention |
|---|---|---|---|
| 0.762 | 0.238 | □ | □ |



Marginal Probabilities

| Drug=no | Drug=yes | Evidence | Intervention |
|---|---|---|---|
| 1.000 | 0.000 | □ | ☑ |

Marginal Probabilities

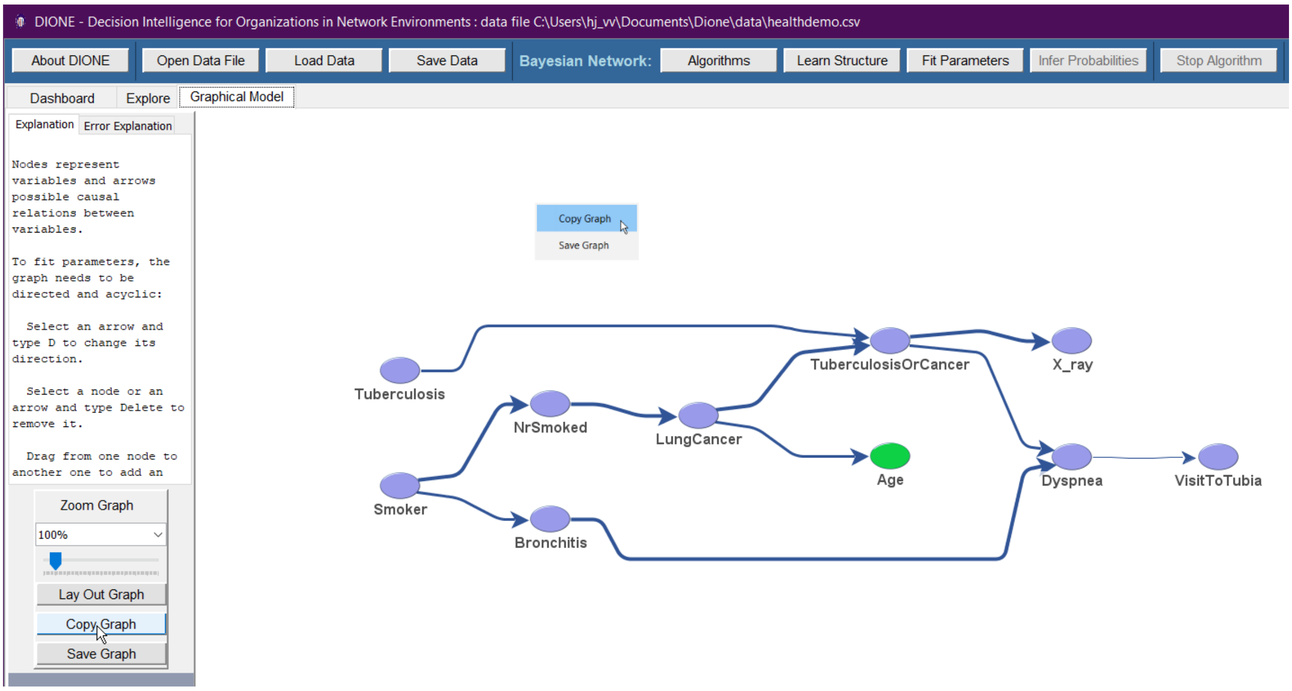
| Heart Attack=no | Heart Attack=yes | Evidence | Intervention |
|---|---|---|---|
| 0.825 | 0.175 | □ | □ |

An intervention means we have given the drug to people regardless of gender, and the inference will eliminate the effect of the confounding variable **Gender**, so the correct conclusion can be drawn that the drug increases the risk of a heart attack for the population.

## Copy and Save Results

You can keep your results by copying the network graph to another application, for example to insert it in a word processing document, or by saving it to a file. After fitting parameters, you can also save numerical results to a text file.

To copy the network graph, press button Copy Graph in the left-hand panel to put a picture of the chart on your computer's clipboard, or, if you prefer, you can right-click on the background of the graph and select Copy Graph from the pop-up menu:



Nodes represent variables and arrows possible causal relations between variables.

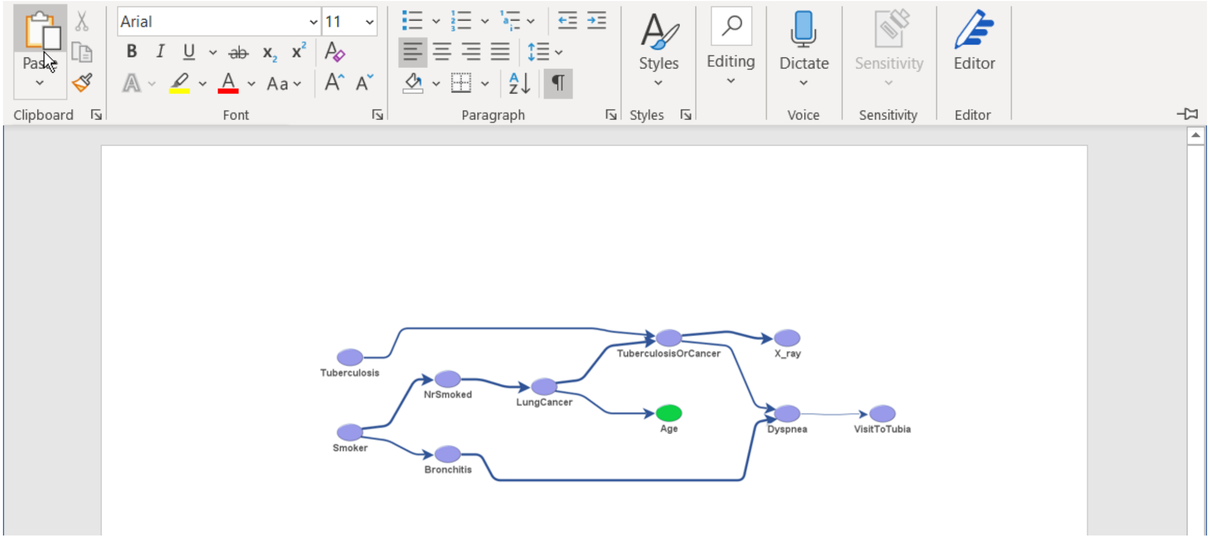
To fit parameters, the graph needs to be directed and acyclic:

Select an arrow and type D to change its direction.

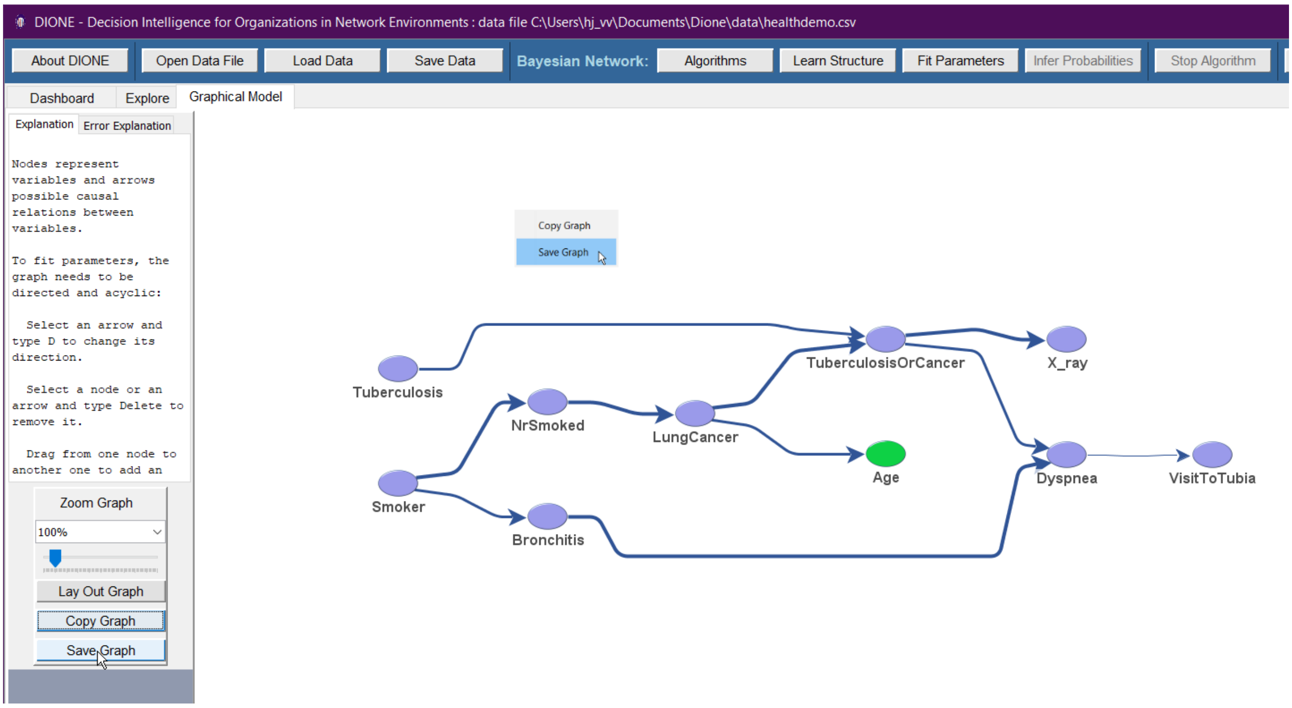
Select a node or an arrow and type Delete to remove it.

Drag from one node to another one to add an

You can now paste the picture in a Word document, for example:



Press button Save Graph in the left-hand panel to save a network graph to a file, or right-click on the graph background and select Save Graph from the pop-up menu:



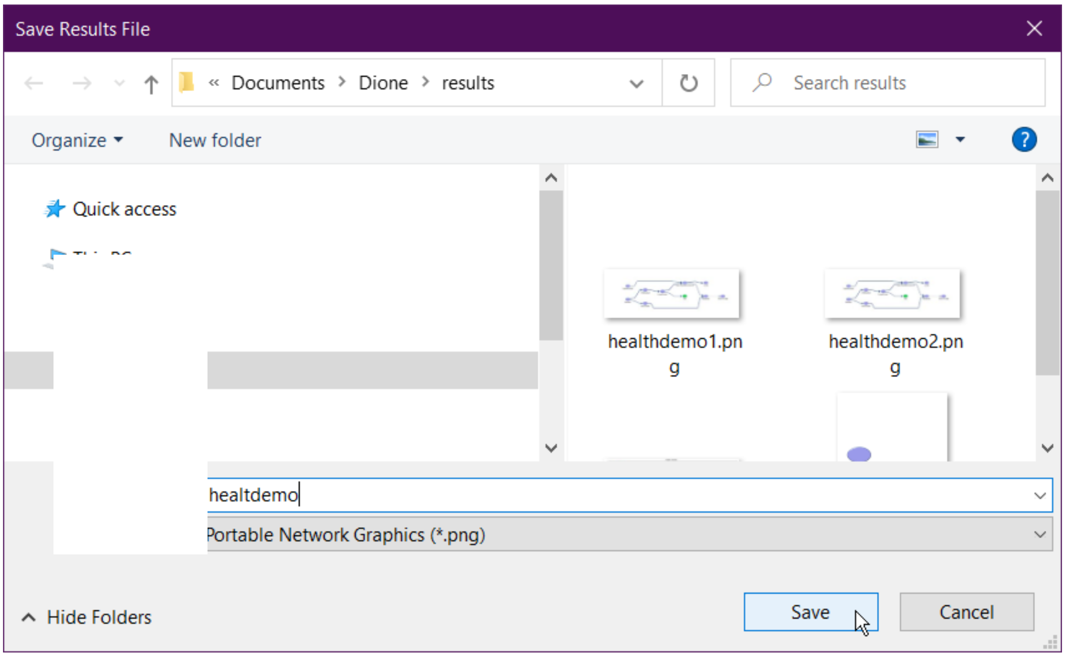
Nodes represent variables and arrows possible causal relations between variables.

To fit parameters, the graph needs to be directed and acyclic:

Select an arrow and type D to change its direction.

Select a node or an arrow and type Delete to remove it.

Drag from one node to another one to add an

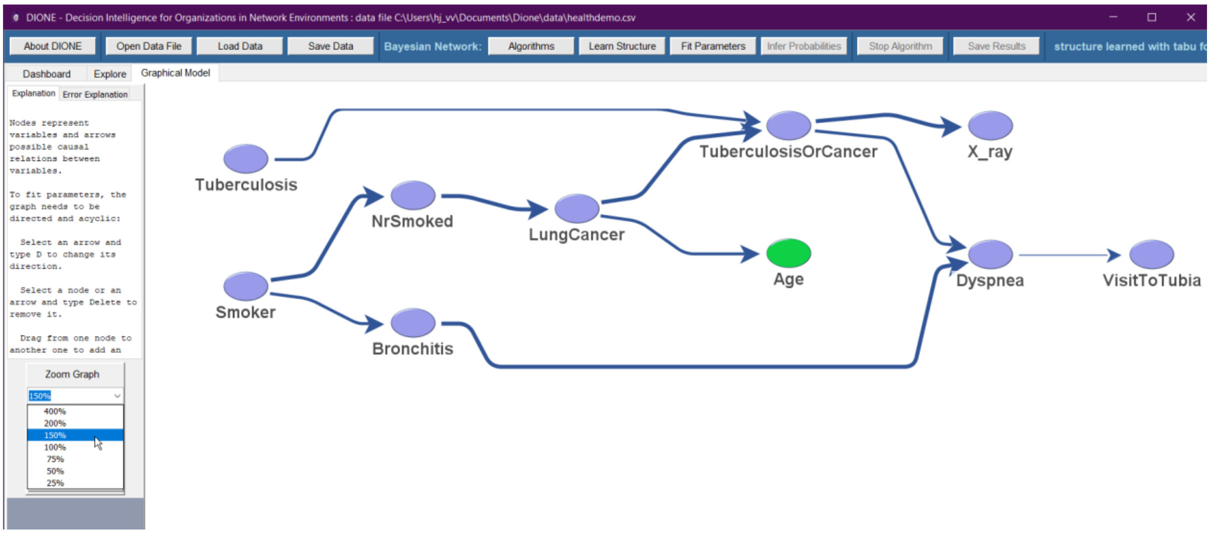


healtdemo

Portable Network Graphics (*.png)

You now have a Portable Network Graphics (PNG) picture file that you van open in any image processing program or insert in a document.

With option Zoom Graph in the left-hand panel you can zoom in or out the network graph:



DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemo.csv

Nodes represent variables and arrows possible causal relations between variables.

To fit parameters, the graph needs to be directed and acyclic:

Select an arrow and type D to change its direction.

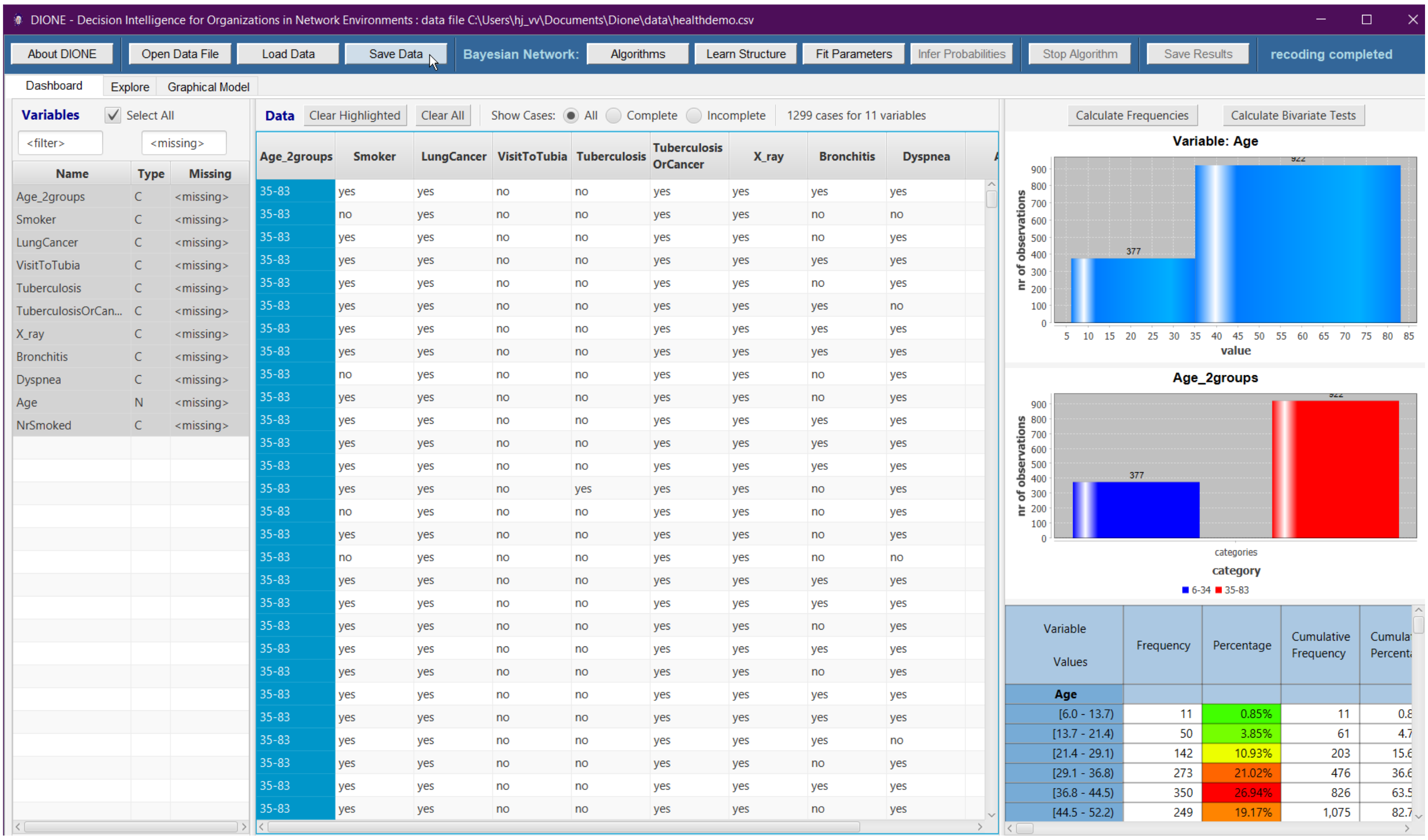
Select a node or an arrow and type Delete to remove it.

Drag from one node to another one to add an

With option Lay Out Graph you can redraw the graph with a nice layout, if it has become messy after a lot of manual editing, for example.

# Save Data

In the current version only data saving to CSV files is supported and in the saved CSV file column headers contain variable names, with values in the columns. The saved data file will have all data of variables shown in the **Variables** table, including any recoded data, for example the recoded **Age** variable here:



DIONE - Decision Intelligence for Organizations in Network Environments : data file C:\Users\hj_vv\Documents\Dione\data\healthdemo.csv

1299 cases for 11 variables

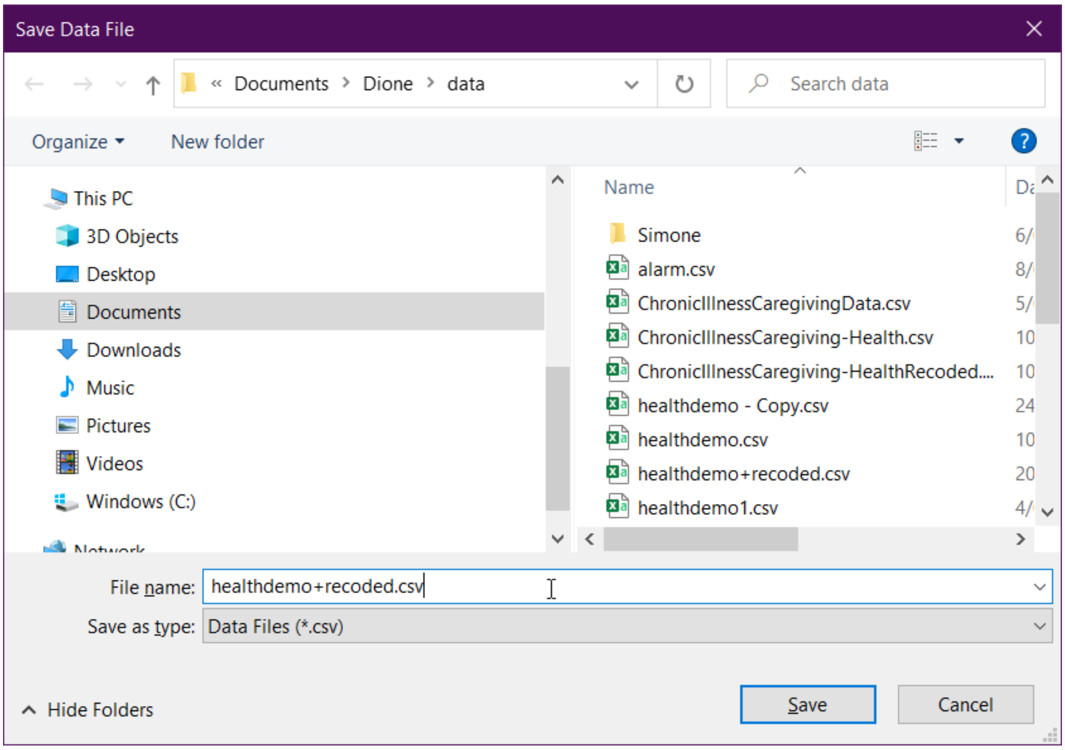
| Name | Type | Missing |
|---|---|---|
| Age_2groups | C | <missing> |
| Smoker | C | <missing> |
| LungCancer | C | <missing> |
| VisitToTubia | C | <missing> |
| Tuberculosis | C | <missing> |
| TuberculosisOrCan... | C | <missing> |
| X_ray | C | <missing> |
| Bronchitis | C | <missing> |
| Dyspnea | C | <missing> |
| Age | N | <missing> |
| NrSmoked | C | <missing> |

| Age_2groups | Smoker | LungCancer | VisitToTubia | Tuberculosis | Tuberculosis OrCancer | X_ray | Bronchitis | Dyspnea |
|---|---|---|---|---|---|---|---|---|
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | no | yes | no | no | yes | yes | no | no |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | no | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | yes | yes | yes | no | yes |
| 35-83 | no | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |
| 35-83 | no | yes | no | no | yes | yes | no | no |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | no |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |
| 35-83 | yes | yes | no | no | yes | yes | yes | yes |
| 35-83 | yes | yes | no | no | yes | yes | no | yes |

| Variable<br>Values | Frequency | Percentage | Cumulative Frequency | Cumula Percent |
|---|---|---|---|---|
| **Age** | | | | |
| [6.0 - 13.7) | 11 | 0.85% | 11 | 0.8 |
| [13.7 - 21.4) | 50 | 3.85% | 61 | 4.7 |
| [21.4 - 29.1) | 142 | 10.93% | 203 | 15.6 |
| [29.1 - 36.8) | 273 | 21.02% | 476 | 36.6 |
| [36.8 - 44.5) | 350 | 26.94% | 826 | 63.5 |
| [44.5 - 52.2) | 249 | 19.17% | 1,075 | 82.7 |

Note that the saved data will not include data for any variables you have removed from the **Variables** table. If radio button ○ Complete or ○ Incomplete has been selected, only complete or incomplete cases will be saved.

To save data to a CSV file, press button Save Data and in the Save Data File dialog enter a name for the data file you wish to create:



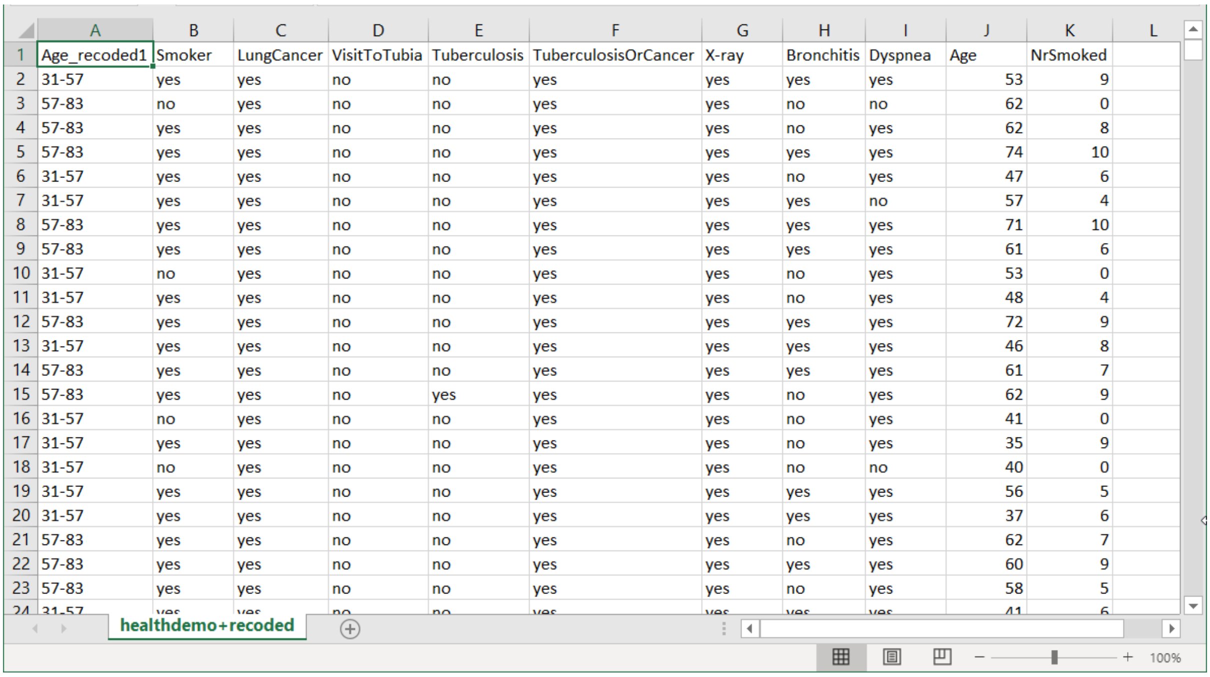
| Name | Da |
|---|---|
| Simone | 6/ |
| alarm.csv | 8/ |
| ChronicIllnessCaregivingData.csv | 5/ |
| ChronicIllnessCaregiving-Health.csv | 10 |
| ChronicIllnessCaregiving-HealthRecoded.... | 10 |
| healthdemo - Copy.csv | 24 |
| healthdemo.csv | 10 |
| healthdemo+recoded.csv | 20 |
| healthdemo1.csv | 4/ |

File name: healthdemo+recoded.csv

Save as type: Data Files (*.csv)

Save Cancel

The saved data file now looks like this:



|  | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age_recoded1 | Smoker | LungCancer | VisitToTubia | Tuberculosis | TuberculosisOrCancer | X-ray | Bronchitis | Dyspnea | Age | NrSmoked |  |
| 2 | 31-57 | yes | yes | no | no | yes | yes | yes | yes | 53 | 9 |  |
| 3 | 57-83 | no | yes | no | no | yes | yes | no | no | 62 | 0 |  |
| 4 | 57-83 | yes | yes | no | no | yes | yes | no | yes | 62 | 8 |  |
| 5 | 57-83 | yes | yes | no | no | yes | yes | yes | yes | 74 | 10 |  |
| 6 | 31-57 | yes | yes | no | no | yes | yes | no | yes | 47 | 6 |  |
| 7 | 31-57 | yes | yes | no | no | yes | yes | yes | no | 57 | 4 |  |
| 8 | 57-83 | yes | yes | no | no | yes | yes | yes | yes | 71 | 10 |  |
| 9 | 57-83 | yes | yes | no | no | yes | yes | yes | yes | 61 | 6 |  |
| 10 | 31-57 | no | yes | no | no | yes | yes | no | yes | 53 | 0 |  |
| 11 | 31-57 | yes | yes | no | no | yes | yes | no | yes | 48 | 4 |  |
| 12 | 57-83 | yes | yes | no | no | yes | yes | yes | yes | 72 | 9 |  |
| 13 | 31-57 | yes | yes | no | no | yes | yes | yes | yes | 46 | 8 |  |
| 14 | 57-83 | yes | yes | no | no | yes | yes | yes | yes | 61 | 7 |  |
| 15 | 57-83 | yes | yes | no | yes | yes | yes | no | yes | 62 | 9 |  |
| 16 | 31-57 | no | yes | no | no | yes | yes | no | yes | 41 | 0 |  |
| 17 | 31-57 | yes | yes | no | no | yes | yes | no | yes | 35 | 9 |  |
| 18 | 31-57 | no | yes | no | no | yes | yes | no | no | 40 | 0 |  |
| 19 | 31-57 | yes | yes | no | no | yes | yes | yes | yes | 56 | 5 |  |
| 20 | 31-57 | yes | yes | no | no | yes | yes | yes | yes | 37 | 6 |  |
| 21 | 57-83 | yes | yes | no | no | yes | yes | no | yes | 62 | 7 |  |
| 22 | 57-83 | yes | yes | no | no | yes | yes | yes | yes | 60 | 9 |  |
| 23 | 57-83 | yes | yes | no | no | yes | yes | no | yes | 58 | 5 |  |
| 24 | 31-57 | yes | yes | no | no | yes | yes | yes | yes | 41 | 6 |  |

# Big Data – Performance

**DIONE** uses efficient R algorithms to load, process and save data, which is quite fast for datasets of up to some 1 million rows of data (also known as cases or records), even on an average laptop PC.

However, when memory and processing power are limited, these tasks become more time-consuming. Some typical times are given here:

| | |
|---|---|
| computer | processor Intel i7, memory 4 GB |
| number of variables | 31 |
| number of cases | 10 000 000 |
| loading data | 2 minutes |
| checking complete cases | 3 minutes |
| number of complete cases | 1 500 000 |
| learning network structure | 17 minutes |
| fitting parameters | 6 minutes |

# References

Judea Pearl & Dana Mackenzie 2018. *The Book of Why – The New Science of Cause and Effect*. Allen Lane, Great Britain.