OECD QSAR Toolbox v.4.1

Example illustrating endpoint vs. endpoint correlation for apical endpoints
Outlook

• Background
• Objectives
• The exercise
• Workflow
Background

This presentation is designed to introduce the user with:

• Illustration of different types endpoint vs. endpoint correlations using:
  ➢ LLNA and GPMT skin sensitization data
  ➢ DPRA and LLNA skin sensitization data
  ➢ Skin sensitization and Ames mutagenicity data
Outlook

• Background
• **Objectives**
• The exercise
• Workflow
Objectives

This presentation demonstrates a number of functionalities of the Toolbox:

• Illustration of endpoint vs. endpoint correlations using different type endpoint data
Outlook

• Background
• Objectives
• **The exercise**
• Workflow
The exercise

• Illustration of different endpoint data correlations:
  - LLNA vs. GPMT skin sensitization data
  - DPRA (reactivity) vs. LLNA (skin sensitization) data
  - GPMT (skin sensitization) vs. Ames mutagenicity data
Outlook

- Background
- Objectives
- The exercise
- **Workflow**
Workflow

• The Toolbox has six modules which are typically used in a workflow:
  • Chemical Input
  • Profiling
  • Endpoints
  • Category Definition
  • Filling Data Gaps
  • Report

• In this example we will use the modules in a different order, tailored to the aims of the example.
Outlook

• Background
• Objectives
• The exercise
• **Workflow**
  • Correlation of data - background
Correlation of endpoint data

Background

• This functionality introduces the user with the opportunity to analyze correlations between selected gap filling endpoint (endpoint used for prediction) and other endpoint data.

• It is applicable for correlation analysis of data presented in ordinary, interval or ratio scale.

• If correlated data are measured in interval or ratio scale they are transformed in ordinary scale and the strength of the correlation is estimated by Spearman correlation coefficient.

• Basically, this functionality provides a correlation between target endpoint (this is the initial endpoint selected by the user) displayed on ordinate axis (Y-axis) and other endpoint data displayed on abscissa (X-axis).
Correlation of endpoint data
Spearman coefficient factor

• Spearman’s rank correlation coefficient is a nonparametric rank statistic proposed by Charles Spearman as a measure of the strength of an association between two variables. It assesses how well the relationship between two variables can be described using a monotonic function.

• Spearman correlation coefficient could be used for exploring the covary between:
  • two ranked variables
  • one measurement variable and one ranked variable (in this case, the measurement variable need to be to converted to ranks)

• Spearman correlation varies from -1 to +1 and the interpretation of the coefficient factor is provided below:
  • 0.00 – 0.19 – very weak correlation
  • 0.20 – 0.39 – weak correlation
  • 0.40 – 0.59 – moderate correlation
  • 0.60 – 0.79 – strong correlation
  • 0.80 – 1.0 – very strong
Outlook

• Background
• Objectives
• The exercise
• **Workflow**
  • Correlation of data – background
  • **Types endpoint correlations**
Types endpoint correlations

Types endpoint correlations are as follows:

- Continuous vs. continuous*
- Categorical vs. categorical:
  - Categorical vs. categorical
  - Categorized continuous vs. categorical
  - Categorized continuous vs. categorized continuous*

*Both type correlation is not illustrated in this presentations. They are presented in “Tutorial_4_TB 4.1_Illustrating endpoint vs. endpoint correlation using ToxCast data”
Outlook

• Background
• Objectives
• The exercise
• **Workflow**
  • Correlation of data – background
  • **Types endpoint correlations**
    • Categorical vs. categorical
The aim of this type correlation is to illustrate how categorical type data correlates each other.

Categorical type data is the statistical data type consisting of categorical variables or of data that has been converted into that form. Such data is binary Ames data (dichotomic type): positive, negative or polytomic type data such as GPMT data: strong, weak and negative.

Two examples illustrating this type correlation will be demonstrated:
- Example 1: Correlation of two types skin sensitization data
  - LLNA (Positive, Negative) vs. GPMT (Weakly positive, Strongly positive, Negative)
- Example 2: Correlation of skin sensitization and Ames mutagenicity data
  - LLNA (Negative, Weakly positive, Strongly positive) vs. AMES (Positive, Equivocal, Negative)

Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
- Load Skin sensitization database (step 1)
- Gather experimental data (step 2)
- Define target endpoint (step 3)
- Enter Gap filling (step 4)
- Perform correlation between endpoints (step 5).
Types endpoint correlations
Categorical vs. categorical
Load Skin sensitization database – step 1

Example 1: Correlation of LLNA and GPMT data

1. Go to “Input”;
2. Click “Database” button;
3. Select “Skin sensitization” database;
4. Click OK;
5. The chemicals from database have been loaded on datamatrix;
Types endpoint correlations
Categorical vs. categorical
Gather experimental data – step 2

Example 1: Correlation of LLNA and GPMT data

1. Go to “Data”; 2. Select “Skin sensitization”; 3. Click “Gather” 4. Click “OK” 5. Click “OK”;

The OECD QSAR Toolbox for Grouping Chemicals into Categories
July, 2017
Types endpoint correlations
Categorical vs. categorical
Gather experimental data – step 2

Example 1: Correlation of LLNA and GPMT data

1. Skin sensitization data appeared on data matrix.
2. Data associated with different type assay (e.g. LLNA, GPMT, HRIPT) are distributed in separate nodes.

All skin sensitization data has been converted into positive/negative data based on implemented scale conversion.

Note: A reminder slide illustrating what is scale and scale conversion is provided on next click.
What is “scale” and “scale conversion”?

Reminder slide

• Skin sensitisation as an example is a “qualitative” endpoint for which the results are presented with categorical type of data (for example: positive; negative; weak sensitizer; strong sensitizer, etc).

• Skin sensitisation potential of the chemicals came from different authors coded with different names (for example: data from John Moores University of Liverpool are: *Strongly sensitizing, Moderately sensitizing etc.*; data from European centre for Ecotoxicology and Toxicology of chemicals are: *Positive, Negative, and Equivocal*).

• The main purpose of the scales is to unify all data available in the Toolbox databases for a certain endpoint.

• “Scale conversion” is the TB instrument to create conversions between scales. More reasonable is to convert more informative to less informative scale.

• The default scale for Skin Sensitisation data is “Skin Sensitisation ECETOC”. It converts all skin sensitization data into: Positive and Negative. This allows skin sensitization data to be used as much as possible for gap filling purposes.
Types endpoint correlations
Categorical vs. categorical
Define target endpoint – step 3

Example 1: Correlation of LLNA and GPMT data

The target endpoint is EC3 data associated with LLNA assay.
1. Click on the cell associated with target endpoint;
Types endpoint correlations
Categorical vs. categorical
*Define target endpoint – step 3*

**Example 1:** Correlation of LLNA and GPMT data

1. **Insert** type “EC3” data associated with LLNA assay in the filter box, then **press** “Enter” and automatically opening the tree to the target endpoint;
2. **Click** on the cell associated with target endpoint;
Types endpoint correlations
Categorical vs. categorical

Enter Gap filling – step 4

**Example 1:** Correlation of LLNA and GPMT data

Follow the steps:
1. **Go** to “Data Gap filling”;
2. **Select** “Read-across”;
3. **Select** “Skin sensitization II (ECETOC)” scale (see Note);
4. **Click** “OK”;

**Note:** By default EC3 data has been converted into binary categories: positive/negative based on scale “Skin sensitization II (ECETOC)”. For the purpose of this exercise, Skin sensitization I (OASIS) will be used. This scale converts EC3 data into three categories: Strongly positive (EC3 0-10%), Weakly positive (EC3 10-50%) and Negative (EC3>50%).

Enter Gap filling and apply read across. Read across is applied because a categorical type data is analyzed.
Types endpoint correlations
Categorical vs. categorical

Enter Gap filling – step 4

Example 1: Correlation of LLNA and GPMT data

The message informing the user for how many chemicals with experimental data are excluded from gap filling due to missing X-descriptor value appeared 1. Click “Ok”;
Types endpoint correlations
Categorical vs. categorical
Perform correlation between LLNA and GPMT data – step 5

Example 1: Correlation of LLNA and GPMT data

Correlation assumes a single value per chemical to be used. In this respect the default calculation settings should be changed from “All” to something different. In our case study we play a worst case scenario, thus an option “All” is changed to “Maximal” values. Follow the steps:
1. **Open** “Calculation options”; 2. **Click** on “Data usage” menu item; 3. **Select** Maximal; 4. **Click** “OK”;
**Types endpoint correlations**

**Categorical vs. categorical**

*Perform correlation between LLNA and GPMT data – step 5*

**Example 1:** Correlation of LLNA and GPMT data

1. Open Descriptor/data tab;
2. Click on Select endpoint tree descriptor;
3. **Open** nodes under “Sensitization” node;
4. **Select** second endpoint, which will be placed on X-axis circled in red box: SMWN;
5. **Click** “OK” button;
6. **Select** Scale I OASIS;
7. **Click** OK
Types endpoint correlations
Categorical vs. categorical
Perform correlation between LLNA and GPMT data– step 5

Example 1: Correlation of LLNA and GPMT data

The message inform the user for how many chemicals will be excluded from correlation due to missing data for SMWN endpoint appears. This will not affect the value of correlation coefficient 1. Click “OK”;

Click “OK”;
Types endpoint correlations
Categorical vs. categorical

Perform correlation between LLNA and GPMT data – step 5

**Example 1:** Correlation of LLNA and GPMT data

1. Select “Descriptor/ data”;
2. Click “Edit descriptor options”;
3. Select “Maximal”;
4. Click OK
Types endpoint correlations
Categorical vs. categorical
Interpretation of correlation results (LLNA vs. GPMT)

Example 1: Correlation of LLNA and GPMT data

- Correlation analysis between two categorical type skin sensitization data (LLNA and GPMT) shows moderate endpoint correlation (Spearman coefficient is 0.54).
Types endpoint correlations
Categorical vs. categorical

- The second example illustrating categorical vs. categorical type correlation is:
  - Example 2: Correlation of Skin sensitization and Ames mutagenicity data
    - LLNA (Negative, Weakly positive, Strongly positive)
    - AMES (Positive, Equivocal, Negative)

- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
  - Load Skin sensitization database (step 1) – skipped, because this database is already loaded on data matrix
  - Gather experimental data (step 2)
  - Define target endpoint (step 3)
  - Enter Gap filling (step 4)
  - Perform correlation between endpoints (step 5)
Types endpoint correlations
Categorical vs. categorical

Gather experimental data – step 2

Sidebar of database relevancy

Once the endpoint is selected, the relevant databases are highlighted.
Types endpoint correlations
Categorical vs. categorical
Gather experimental data – step 2

Example 2: Correlation of LLNA and AMES data

1. In order to start with next example please **click** on the level Skin sensitization from document tree
2. **Click** on Data tab
Types endpoint correlations
Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation of LLNA and AMES data

1. Remove EC3 from the filter, click Enter
2. Position on the level of genetic toxicity as shown
Types endpoint correlations
Categorical vs. categorical
Gather experimental data – step 2

Example 2: Correlation of LLNA and AMES data

1. Select the databases including Ames data (green highlighted). Do not check ECHA Chem database.
2. Skin sensitization DB is already selected;
3. Click “Gather”;
4. The data appeared on datamatrix;

Note that the correlation between endpoints is possible when data is gathered and available on data matrix. One should be aware of the data values that would be using during the data gap filling and gather the data for the corresponding endpoint during the “Endpoint” stage of the workflow, prior to entering the “Data gap filling” module.
Types endpoint correlations
Categorical vs. categorical
*Define target endpoint – step 3*

**Example 2:** Correlation of LLNA and AMES data

The target endpoint is skin sensitization/in vivo/LLNA/EC3;
1. **Click** on the cell associated with target endpoint;
Types endpoint correlations
Categorical vs. categorical

Enter Gap filling – step 4

Example 2: Correlation of LLNA and AMES data

Enter Gap filling applying read across. Read across is applied because a categorical type data is analyzed.
1. Go to "Data Gap filling"; 2. Select "Read-across"; 3. Check "Skin sensitization I (OASIS)" scale; 4. Click "OK"
Example 2: Correlation of LLNA and AMES data

The message informs the user for chemicals excluded from gap filling. 1. Click “OK”;
Types endpoint correlations
Categorical vs. categorical
Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of LLNA and AMES data

1. Open “Calculation” options.
2. Click on “Data usage”
3. Select “Maximal”
4. Click “OK” (refer to slide 62 for details)
Types endpoint correlations
Categorical vs. categorical
Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of LLNA and AMES data

1. Open Select/descriptors data/ Select endpoint tree descriptor; 2. Open nodes under “Genetic Toxicity” node; 3. Select “With S9” under In Vitro|Bacterial Reverse Mutation Assay (e.g. Ames Test)|Gene Mutation| Salmonella typhimurium; 4. Click “OK” button;
Types endpoint correlations
Categorical vs. categorical
Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of LLNA and AMES data

Possible data inconsistency window is appearing. Click OK.
Types endpoint correlations
Categorical vs. categorical
Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of LLNA and AMES data

The appearing message inform for the common number gathered data across the number chemicals that will be excluded in Trend analysis due to missing X descriptor value(s). They are analogues with no AMES data. This will not affect the value of correlation coefficient; 1.

Click “OK”;

The OECD QSAR Toolbox for Grouping Chemicals into Categories
July, 2017
Types endpoint correlations
Categorical vs. categorical
Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of GPMT and AMES data

1. Select “Descriptor/ data”; 2. Click “Edit descriptor options”; 3. Select “Maximal” (worst case); 4. Click OK
Correlation analysis between two categorical type data: GPMT and AMES shows weak correlation between two endpoints (Spearman coefficient is 0.3).
Outlook

• Background
• Objectives
• The exercise
• **Workflow**
  • Correlation of data – background
  • **Types endpoint correlations**
    • Categorical vs. categorical
    • Categorized continuous vs. categorical
Types endpoint correlations
Categorized continuous vs. categorical

• The aim of this type correlation is to illustrate how categorized continuous and categorical type data correlates each other.

• Categorized continuous data is the continuous type data (e.g. LC50 or AC50, EC3, %) converted into categories.

• In this example we will illustrated how DPRA ratio data (%) correlates with LLNA data:
  • DPRA (ratio data expressed in % and converted in categories)
  • LLNA (categorical type: Strongly positive, Weakly positive, Negative)

• Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
  • Load Skin sensitization database (step 1) – skipped, because this database has been already loaded on data matrix
  • Gather experimental data (step 2)
  • Define target endpoint (step 3)
  • Enter Gap filling (step 4)
  • Perform correlation between endpoints (step 5).
Types endpoint correlations
Categorized continuous vs. categorical

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

1. Click again on the level of Skin sensitization; 2. Position the mouse on the level of In Chemico level of endpoint tree; 3. Click on Data tab
Types endpoint correlations
Categorized continuous vs. categorical

*Example:* Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

1. Go to “Data”;
2. Select “Chemical reactivity COLIPA” database. Skin sensitization DB is already selected;
3. Click “Gather” button;
4. The data appeared on datamatrix;

The OECD QSAR Toolbox for Grouping Chemicals into Categories

July, 2017
Types endpoint correlations
Categorized continuous vs. categorical
Define target endpoint – step 3

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

The target endpoint is EC3; 1. **Click** on the cell associated with target endpoint and target chemical;
Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

Enter Gap filling and apply read across. Read across is applied because a categorical type data is analyzed.

1. Go to “Data Gap filling”;
2. Select “Read-across”;
3. Select “Skin sensitization I (OASIS)” scale;
4. Click “OK”;
5. Enter Gap filling – step 4

Original data

Converted data
Types endpoint correlations
Categorized continuous vs. categorical
Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data
Types endpoint correlations
Categorized continuous vs. categorical
Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

1. Open Descriptors/data of “DPRA”. In this case we mixed DPRA lysine and Cysteine data
2. Click on “Select endpoint tree descriptor” button
3. Click on OK on the appeared message
4. Click on OK
5. Select Chemical reactivity DPRA 13% (ordinal) scale
6. Click OK
7. Click OK on the appeared message
Types endpoint correlations
Categorized continuous vs. categorical
Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

1. Open Edit descriptor options
2. Select maximal values (worst case)
3. Click OK
Types endpoint correlations
Categorized continuous vs. categorical

Interpretation of correlation results (DPRA vs. LLNA)

- In this example we have correlate continues DPRA (%) data distributed into 3 bins (shown below) and categorical LLNA data (Strongly positive, Weakly positive, Negative)
  - Less than 9%
  - Grey zone 9 – 21%
  - Above 21%

- The high value of Spearman coefficient (0.49) shows moderate correlation between DPRA and LLNA data
Summary

- Different type correlations have been illustrated in this tutorial based on type of endpoint data:
  - Categorical vs. categorical:
  - Categorized continuous vs. categorical

- Correlation analysis has been evaluated by Spearman coefficient

- Moderate endpoint correlations have been obtained for 2 out of 3 illustrated examples.