



www.emif.eu

European Medical Information Framework

Grant Agreement n°115372

D13.2 Data analysis tools for vertical projects version 1

WP13 – Analysis, processing & visualization methods and tools

V1.1

Final

Lead beneficiary: JANSSEN

Date: 26/03/2015

Nature: P

Dissemination level: PU

Reproduction of this document or part of this document without EMIF consortium permission is forbidden. Any use of any part must acknowledge the EMIF consortium as “EMIF European Medical Information Framework, grant agreement n° 115372 (Innovative Medicines Initiative Joint Undertaking)”. This document is shared in the EMIF Consortium under the conditions described in the EMIF Project Agreement, section 8.



 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.1 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	2/21

TABLE OF CONTENTS

DOCUMENT INFORMATION	3
DOCUMENT HISTORY.....	4
DEFINITIONS	5
ABBREVIATIONS	7
EXECUTIVE SUMMARY.....	8
KEY WORDS (WORDLE STYLE)	9
1. INTRODUCTION.....	10
2. DATASET EXPLORER.....	11
2.1. ADVANCED ANALYSES AND VISUALIZATIONS	12
2.1.1. <i>Example 1: Survival Analysis</i>	13
2.1.2. <i>Example2: Correlation analysis</i>	15
2.1.3. <i>Example3: Box Plot with Anova</i>	16
2.2. EXPORTING DATA	17
3. FRAMEWORK FOR R PLUGINS	19
4. CROSS TRIAL QUERY	19
4.1. CROSS TRIAL QUERIES IN TRANSMART	19
4.1.1. <i>Example 1: Compare two cohorts using Summary statistics</i>	20
4.1.2. <i>Example 2: Compare clinical rating scale results between two cohorts</i>	20
4.1.3. <i>Example 3: Linegraph showing time course data per cohort</i>	21
4.2. DATA LOAD FOR CROSS TRIAL QUERIES	21

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.1 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	3/21


DOCUMENT INFORMATION

Grant Agreement Number	115372	Acronym	EMIF
Full title	European Medical Information Framework		
Project URL	http://www.emif.eu		
IMI Project officer	Ann Martin (Ann.Martin@imi.europa.eu)		

Deliverable	Number	13.2	Title	Data analysis tools for vertical projects version 1
Work package	Number	13	Title	Analysis, processing & visualization methods and tools


Delivery date	Contractual	Month 24	Actual	26/03/2015
Status	V1.1		Draft <input type="checkbox"/> Final <input checked="" type="checkbox"/>	
Nature	Report <input type="checkbox"/> Prototype <input checked="" type="checkbox"/> Other <input type="checkbox"/>			
Dissemination Level	Public <input checked="" type="checkbox"/> Restricted <input type="checkbox"/> Confidential <input type="checkbox"/>			

Authors (Partner)	Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)		
Responsible Author	Rudi Verbeeck	Email	rverbeec@its.jnj.com
	Partner JANSSEN	Phone	+32 14 641034

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.1 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	4/21


DOCUMENT HISTORY

NAME	DATE	VERSION	DESCRIPTION
Rudi Verbeeck	17-nov-2014	0.1	Table of contents and section descriptions
Janneke Schoots – van der Ploeg	17-nov-2014	0.11	Draft content Dataset Explorer, Framework for R plugins and Cross trial query sections
Rudi Verbeeck	1-dec-2014	0.2	Executive summary and introduction
Janneke Schoots – van der Ploeg	5-jan-2015	0.3	Added description for correlation analysis and boxplot anova
Rudi Verbeeck	12-jan-2015	0.4	Version for internal EBI review
Natalja Kurbatova, Alvis Brazma	20-jan-2015	0.5	Version reviewed by EBI
Janneke Schoots – van der Ploeg	06-feb-2015	0.6	Accepted changes by Natalja and Alvis, and modified Cross trial query section
Rudi Verbeeck	13-feb-2015	0.7	Formatting. Version for internal review.
Janneke Schoots – van der Ploeg, Rudi Verbeeck	25-mar-2015	1.0	Corrections and additions after internal review. Final version.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.1 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	5/21


DEFINITIONS

- Partners of the EMIF Consortium are referred to herein according to the following codes:
 - **Janssen.** Janssen Pharmaceutica NV (Belgium) - **Coordinator**
 - **EMC.** Erasmus University Rotterdam (Netherlands) - **Managing entity of the IMI JU funding**
 - **SYNAPSE.** Synapse Research Management Partners (Spain)
 - **UCL.** University College London (United Kingdom)
 - **PENTA.** Fondazione PENTA (Italy)
 - **EMBL.** European Molecular Biology Laboratory (Germany)
 - **EuroRec.** The European Institute for Health Records (France)
 - **UAVR.** Universidade de Aveiro (Portugal)
 - **ULEIC.** University of Leicester (United Kingdom)
 - **UPF.** Universitat Pompeu Fabra (Spain)
 - **APHP-HEGP.** Assistance Publique - Hôpitaux de Paris (France)
 - **ARS.** Agenzia Regionale di Sanità (Italy)
 - **PHARMO.** PHARMO Institute N.V. (Netherlands)
 - **AUH.** Aarhus Universitetshospital (Denmark)
 - **GENOMEDICS.** Genomedics S.R.L (Italy)
 - **BIPS.** Leibniz Institute for Prevention Research and Epidemiology – BIPS GmbH (Germany)
 - **PEDIANET.** Societa Servizi Telematici SRL (Italy)
 - **UTARTU.** University of Tartu (Estonia)
 - **UNIMAN.** University of Manchester (United Kingdom)
 - **BF.** Brighton Collaboration Foundation (Switzerland)
 - **CUSTODIX.** Custodix NV (Belgium)
 - **UGOT.** University of Gothenburg (Sweden)
 - **KI.** Karolinska Institute (Sweden)
 - **UCAM.** University of Cambridge (United Kingdom)
 - **UH.** University of Helsinki (Finland)
 - **UEF.** University of Eastern Finland (Finland)
 - **UNIPI.** University of Pisa (Italy)
 - **INSERM.** Institute National de la Santé et de la Recherche Médicale (France)
 - **UPS.** Université Paul Sabatier Toulouse III (France)
 - **VKJK.** Vestische Kinder-und Jugendklinik (Germany)
 - **UOL.** University of Leipzig (Germany)
 - **UGLA.** University of Glasgow (United Kingdom)
 - **UCPH.** University of Copenhagen (Denmark)
 - **GSK.** GlaxoSmithKline Research & Development Limited (United Kingdom)
 - **PFIZER.** Pfizer Limited (United Kingdom)
 - **NOVO.** Novo Nordisk (Denmark)
 - **ROCHE.** F. Hoffmann-La Roche AG (Switzerland)
 - **SERVIER.** Servier (France)
 - **Boehringer Ingelheim.** Boehringer Ingelheim International GmbH (Germany)
 - **AMGEN.** Amgen (Belgium)
 - **UCB.** UCB Pharma SA (Belgium)

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.1 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	6/21


- **KCL.** King's College London (United Kingdom). **Co-Coordinator**
- **VUMC.** VU Medical Centre (Netherlands)
- **Concentris.** Concentris research management GmbH (Germany)
- **IRCCS.** IRCCS-FBF (Italy)
- **UPMC.** Université Pierre et Marie Curie (France)
- **UA.** Universiteit Antwerpen (Belgium)
- **UKER.** University of Erlangen (Germany)
- **UM.** University of Maastricht (Netherlands)
- **AE.** Alzheimer Europe (Luxembourg)
- **PSPLC.** Electrophoretics Ltd (United Kingdom)
- **VIB.** VIB University of Antwerp (Belgium)
- **MAAT.** MAAT (France)
- **EHNT.** Ealing Hospital NHS Trust (United Kingdom)
- **CAMCOG.** Cambridge Cognition Ltd (United Kingdom)
- **UOXF.** University of Oxford (United Kingdom)
- **MERCK.** Merck (Germany)
- **UZL.** Universität zu Lübeck (Germany)

- **Grant Agreement.** The agreement signed between the beneficiaries and the IMI JU for the undertaking of the EMIF project (115372).
- **Project.** The sum of all activities carried out in the framework of the Grant Agreement.
- **Topic.** The sum of all activities carried out in the framework of either the EMIF-Platform, the EMIF-AD or the EMIF-Metabolic subprojects within the EMIF Project.
- **Work plan.** Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out, as specified in Annex I to the Grant Agreement.
- **Consortium.** The EMIF Consortium, comprising the above-mentioned legal entities.
- **Project Agreement.** Agreement concluded amongst EMIF participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.1 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	7/21

ABBREVIATIONS

- **MMSE.** Mini-mental state examination, a commonly used questionnaire to assess cognitive function.
- **ANOVA.** Analysis of variance, a statistical method in which the variation in a set of observations is divided into distinct components. One-way ANOVA is used to test for differences among two or more independent groups (means). An ANOVA hypothesis tests the difference in population means (continuous dependent variable) based on one characteristic or factor (categorical independent variable). Two-way analysis of variance (ANOVA) is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable. Multivariate analysis of variance (MANOVA or MNOVA) is used when there are several continuous dependent variables. MNOVA tests for the difference in two or more vectors of means.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.1 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	8/21

EXECUTIVE SUMMARY


Deliverable 13.2 “Data analysis tools for vertical projects version 1” details the functionality in the tranSMART application that is relevant to the analysis and visualization of clinical cohort data in EMIF.

An interactive data analysis and visualization environment such as tranSMART needs to balance usability and general functionality against specialized methods. General exploratory functions, including descriptive univariate statistics, are provided in the tranSMART dataset explorer. Commonly used statistical methods are available in the Advanced Statistics tab.

Specialized methods required to solve a specific research question can be plugged into tranSMART if warranted by usage or demand. Generally though, and in line with expectations from the EMIF verticals, this type of analysis is performed outside of tranSMART in custom tools, such as R. Data export functionality should therefore be available, but should also enforce security restrictions on the data as specified by the data owners.

The requirement from EMIF AD to be able to query variables across trials, as specified in Use Case 1 (Deliverable 9.1 “Initial requirements and benchmarks set”) expects special attention to data loading procedures. Methods to ensure data is harmonized appropriately are outside the scope of this deliverable (see deliverables 11.3 “Extended Specification of the Framework of Reference” and 14.5 “A data management solution for vertical projects, version 2” for more details). Here we concentrate on the required extra steps in the data upload process.

The current deliverable describes the tranSMART Dataset Explorer, Advanced Statistics, R Plugin Framework, Data Export and Cross Trial Query functionality.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.0 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	10/21

1. INTRODUCTION

The first version of the EMIF data analysis and visualization environment concentrates on the analysis of clinical variables.

A typical data analysis project follows a path through the data landscape from a high level overview of available variables over univariate statistics to a gradually deeper investigation of measurements relevant to the research question. The process of selecting relevant variables and cohorts for a particular research question is supported by the EMIF Catalogue and is covered in deliverable 14.3 “EMIF-Platform, version 1”. This deliverable concentrates on the analysis functionality available in tranSMART that covers the next steps in the process.


Section 2 (DATASET EXPLORER) explains the functionality of the tranSMART dataset explorer to view available studies and variables, and to generate summary statistics. The dataset explorer is typically used as an entry point in tranSMART to start an analysis project.

Specialized analyses that support specific use cases are listed in section 2.1 (Advanced Analyses and Visualizations). Depending on the scientific question, specific statistical methods can be applied to investigate the data in more detail.

A general data analysis and visualization environment cannot provide for every possible statistical calculation. TranSMART advanced analysis methods focus on the standard statistical algorithms. Custom methods can be applied in a two-step process:

- **Data export:** subject to approval by the data owner(s), data can be exported from tranSMART for upload and analysis in custom statistical applications such as R. The users are responsible for the statistical methods and visualizations used in the analysis scripts. Data export from tranSMART is covered in section 2.2 (Exporting Data).
- **Plugins:** when an analysis method gets accepted as a standard and is routinely used it can be made available in tranSMART as a plugin. TranSMART currently supports plugins developed in R, but requires some additional development in groovy/grails for the data input page and the page displaying the results of the R calculations and graphs. Section 3 (Framework for R plugins) explains the steps required to develop an R plugin.

A user requirement from EMIF AD that we have paid particular attention to is the pooling (merging) of data for cross trial query. The current approach to data harmonization is detailed in deliverable 14.5 (“A data management solution for vertical projects, version 2”) and a description of methods and semantic web technologies for a more sustainable approach are provided in deliverable 11.3 “Extended Specification of the Framework of Reference”. TranSMART version 1.2 has some support for cross trial functionality but requires additional information to be provided during the data loading process. Section 4 (Cross Trial Query) explains the required steps.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	11/21

2. DATASET EXPLORER

The Dataset Explorer part (called ‘Analyze’ in the newly released TranSMART version 1.2 interface) allows to compare data and test hypotheses for subjects in two different study groups, based on specified criteria and points of comparison.

A user can select the study of interest using drag and drop functionality from a navigation tree and select the points of comparison between two study groups. TranSMART provides summary data about the subjects being compared, and several different views of the comparison data, see Figure 1 - Figure 3.

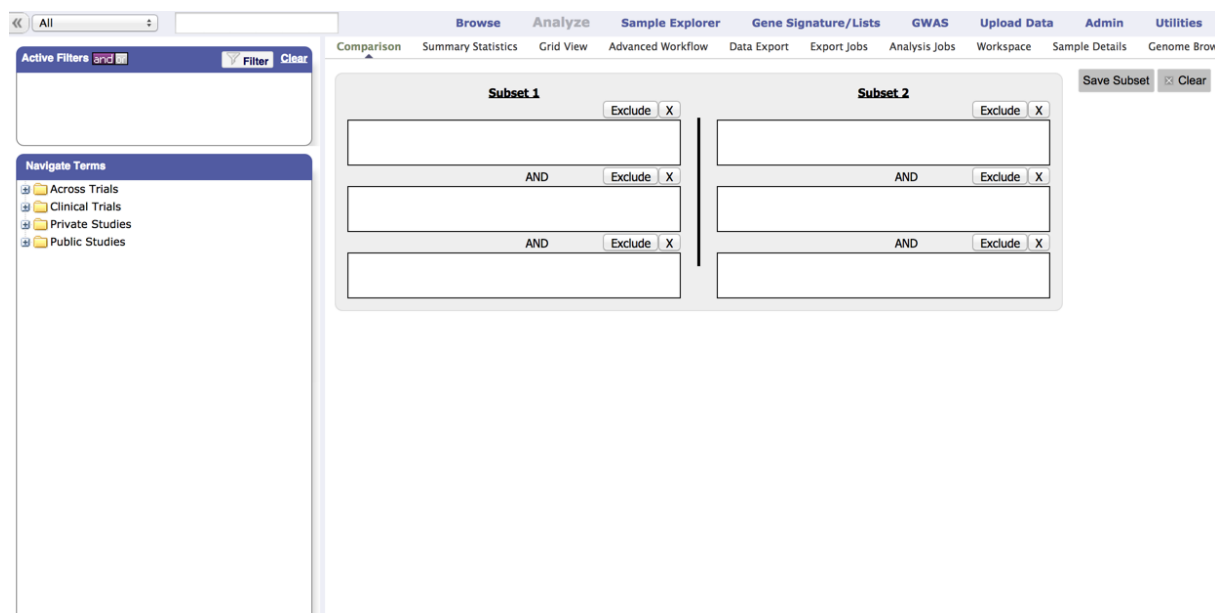



Figure 1. TranSMART dataset explorer.

 IMi - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	12/21

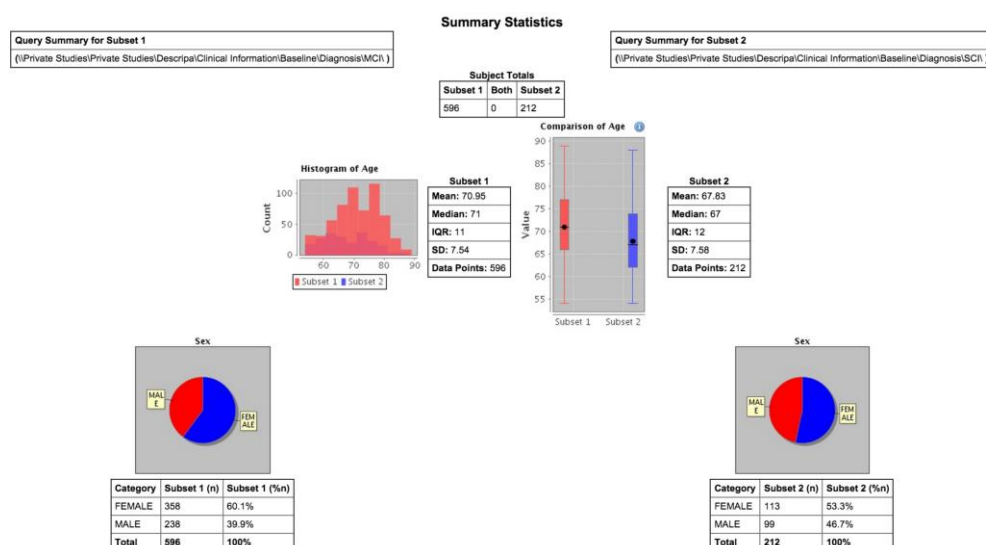


Figure 2. TranSMART Summary Statistics, comparing subjects with MCI (mild cognitive impaired) and SCI (subjective cognitive impaired).

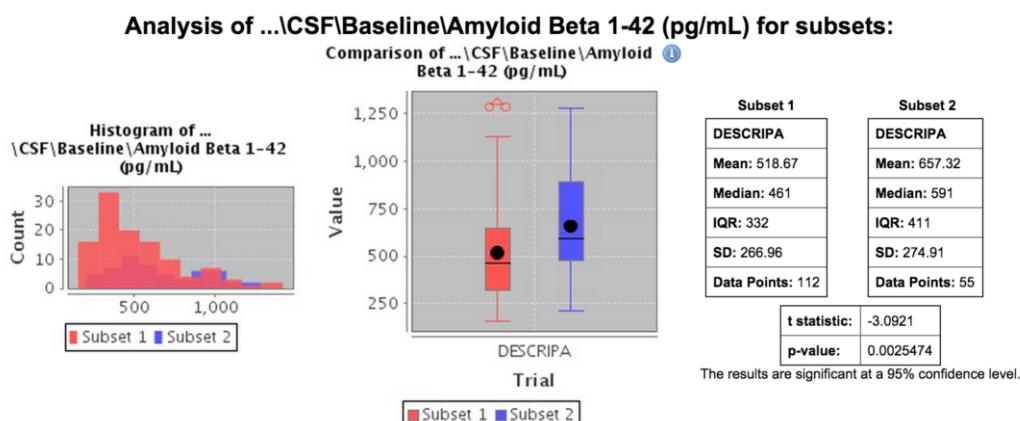



Figure 3. TranSMART Summary Statistics, comparison of CSF Amyloid Beta 1-42 (pg/mL) for MCI and SCI subjects.

2.1. Advanced Analyses and Visualizations

Advanced analyses and visualizations offered with tranSMART allow a user to run the following analyses within the Dataset Explorer:

- Box Plot with ANOVA
- Principal Component Analysis
- Scatter Plot with Linear Regression
- Survival Analysis (See below)
- Table with Fisher Test Analysis
- Standard Heatmap

 IMM - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	13/21

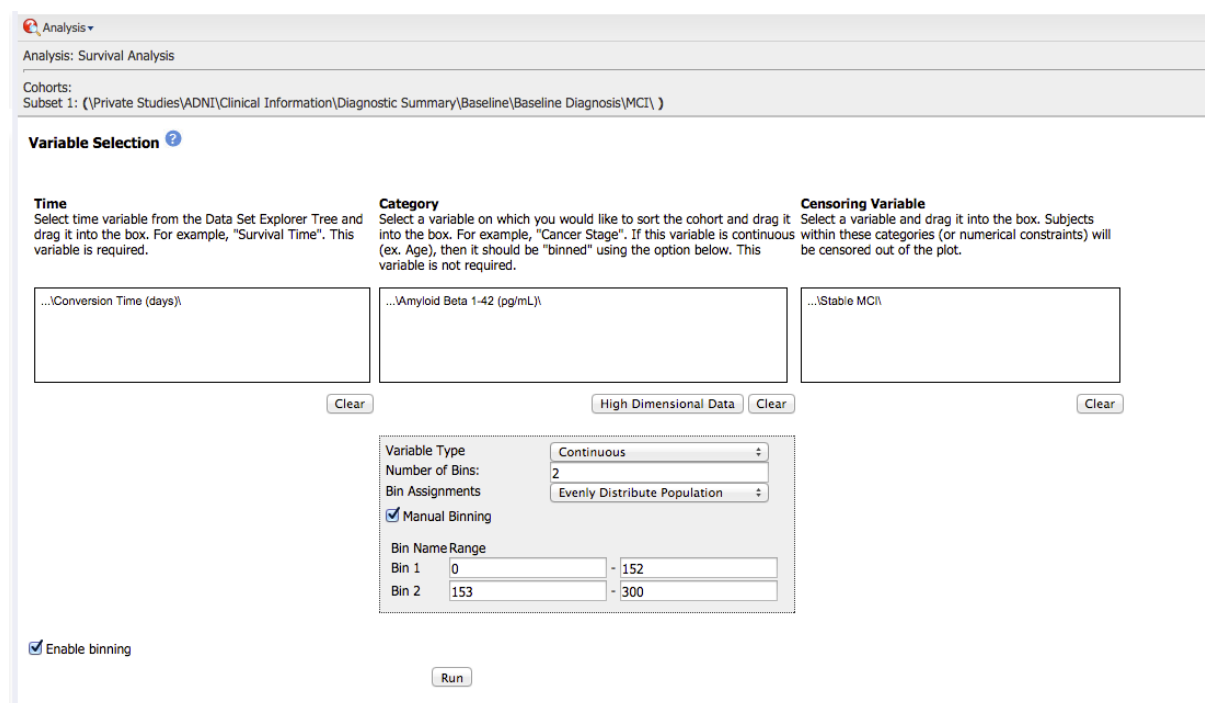
- Hierarchical Clustering
- K-Means Clustering
- Marker Selection (top differentially expressed markers between 2 subpopulations).
- ... and others

Dataset Explorer uses the R software environment for statistical computing to generate analyses and visualizations.

2.1.1. Example 1: Survival Analysis

A survival analysis displays time-to-event data. After defining a cohort of subjects the Variable Selection section appears. To perform a survival analysis three more variables have to be defined (Figure 4):

- Time: a variable that indicates time to conversion
- Category: variable that defines groups into which the data will be split in order to compare their survival times. Continuous variables need to be binned.
- Censoring: variable that selects subjects that are censored out of the plot).



Analysis: Survival Analysis

Cohorts:
Subset 1: (\\Private Studies\\ADNI\\Clinical Information\\Diagnostic Summary\\Baseline\\Baseline Diagnosis\\MCI\\)

Variable Selection ?

Time
Select time variable from the Data Set Explorer Tree and drag it into the box. For example, "Survival Time". This variable is required.

...\\Conversion Time (days)\\

Category
Select a variable on which you would like to sort the cohort and drag it into the box. For example, "Cancer Stage". If this variable is continuous (ex. Age), then it should be "binned" using the option below. This variable is not required.

...\\Amyloid Beta 1-42 (pg/mL)\\

Censoring Variable
Select a variable and drag it into the box. Subjects within these categories (or numerical constraints) will be censored out of the plot.

...\\Stable MCI\\

Clear High Dimensional Data Clear


Variable Type: Continuous
Number of Bins: 2
Bin Assignments: Evenly Distribute Population
☒ Manual Binning

Bin Name	Range
Bin 1	0 - 152
Bin 2	153 - 300

☒ Enable binning

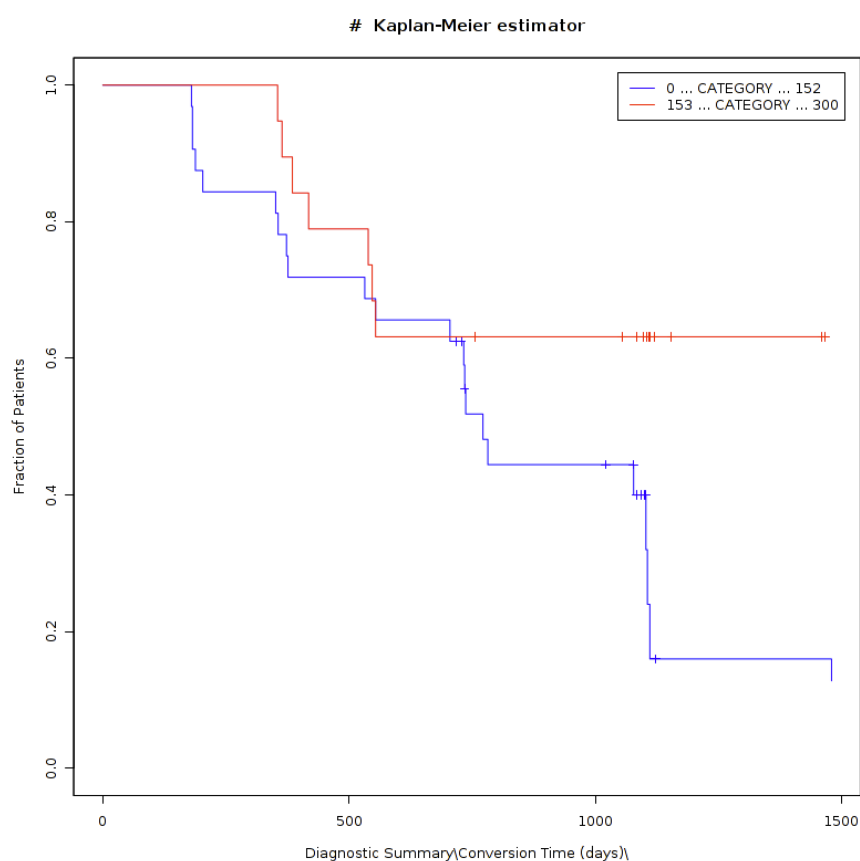
Run

Figure 4. TransSMART Survival Analysis input form.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)		Security: PU 14/21

The analysis results are presented in the form of a graph and statistical summary output in the table format (Figure 5).

Survival Curve



Cox Regression Result


Number of Subjects	51
Number of Events	29
Likelihood ratio test	3.86 on 1 df, p=0.04949
Wald test	3.44 on 1 df, p=0.06349
Score (logrank) test	3.63 on 1 df, p=0.05677

Subset	Cox Coefficient	Hazards Ratio	Lower Range of Hazards Ratio, 95% Confidence Interval	Upper Range of Hazards Ratio, 95% Confidence Interval
153 \342\211\244 CATEGORY \342\211\244 300	-0.8167	0.4419	0.1865	1.047

Survival Curve Fitting Summary

Subset	Number of Subjects	Max Subjects	Subjects at Start	Number of Events	Median Time Value	Lower Range of Time Variable, 95% Confidence Interval	Upper Range of Time Variable
0 \342\211\244 CATEGORY \342\211\244 152	32	32	32	22	772	705	NA
153 \342\211\244 CATEGORY \342\211\244 300	19	19	19	7	NA	554	NA

Figure 5. tranSMART Survival Analysis output.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	15/21

2.1.2. Example2: Correlation analysis

A correlation analysis calculates measure of statistical dependence between two variables and displays graphical output. When the cohorts for analysis are defined in order to perform a correlation analysis, the user has to define variables to analyse and the correlation type. The following correlation types are supported: Spearman correlation is the default type and assess how well the relationship between two variables can be described by using a monotonic function; other correlation types supported are Pearson and Kendall.

The input screen and Spearman correlation analysis results are shown in Figure 6.

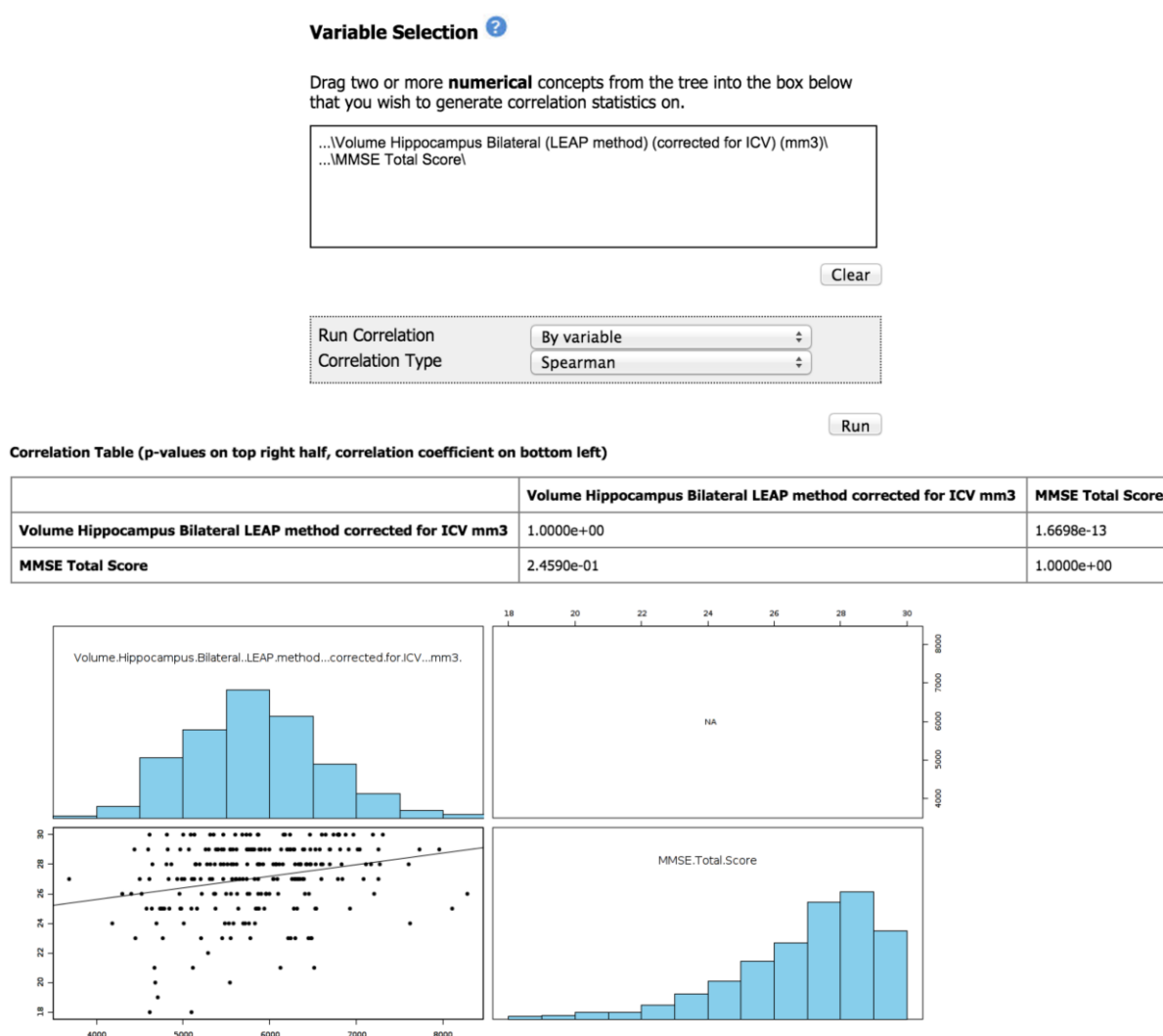



Figure 6. Correlation analysis in TranSMART: input form & results page.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.0 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	16/21

2.1.3. Example3: Box Plot with Anova

Analysis of variance (ANOVA) can be performed on the defined variables from the selected cohort. One-way ANOVA, two-way ANOVA or MNOVA is performed depending on the selection of dependent and independent variables. For example: Independent variable is Hippocampal volume, dependent variable is ApoE4 Genotype (see Figure 7).

Variable Selection

Independent Variable

Select a variable from the Data Set Explorer Tree and drag it into the box. At least one of the variables selected should be a continuous variable (ex. Age) and one should be a categorical variable (ex. Tumor Stage). A continuous variable can be categorized using the binning option below.

...\Volume Hippocampus Bilateral (LEAP method) (corrected for ICV) (mm3)\

Dependent Variable

Select a variable from the Data Set Explorer Tree and drag it into the box. At least one of the variables selected should be a continuous variable (ex. Age) and one should be a categorical variable (ex. Tumor Stage). A continuous variable can be categorized using the binning option below.

...\Heterozygote\
...\Homozygote\
...\Non-carrier\

High Dimensional Data Clear


High Dimensional Data Clear

☐ Enable binning

Run

Figure 7. TranSMART Box plot with ANOVA input form.

The two-way ANOVA results in the form of a boxplot and tables are shown in Figure 8.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	17/21

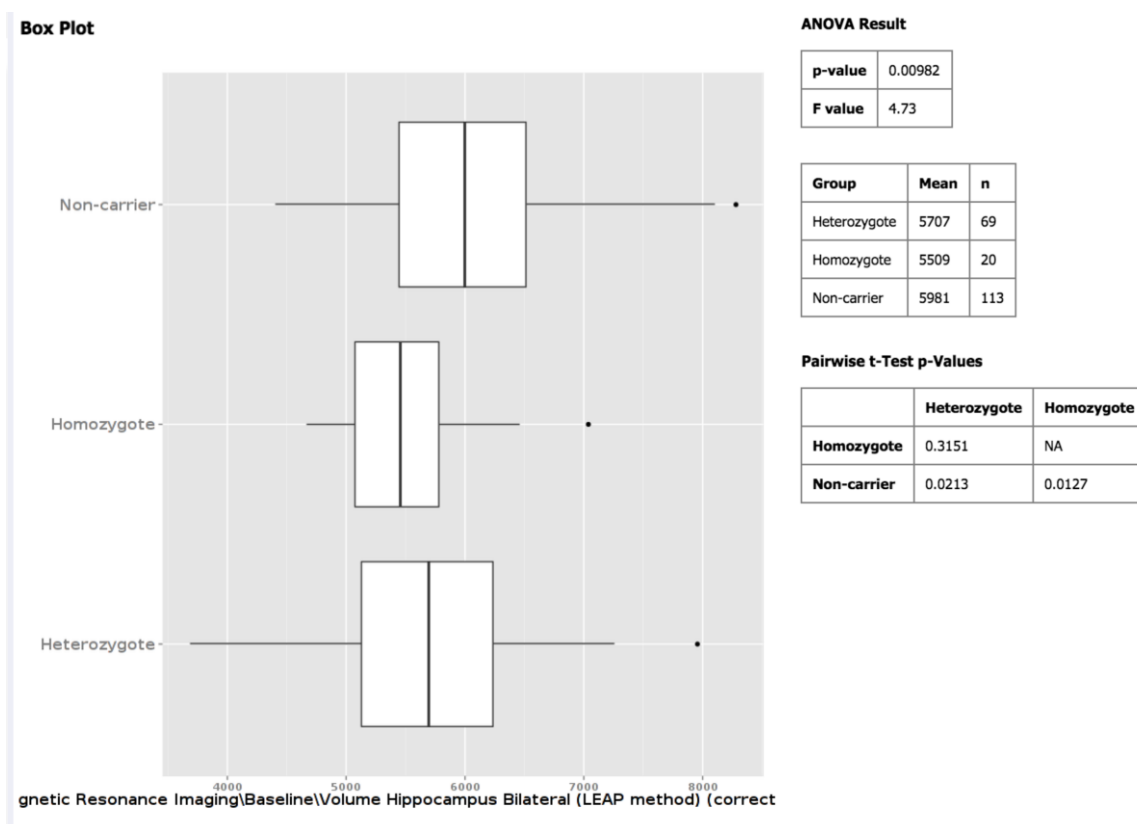



Figure 8. TranSMART Box plot with ANOVA sample output.

2.2.Exporting Data

The Data Export functionality of tranSMART allows exporting data locally for further analysis in several different formats. After the cohorts of interest selection, user can choose the format of data export (see Figure 9 for Data Export menu and Figure 10 for format selection menu examples).

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	18/21

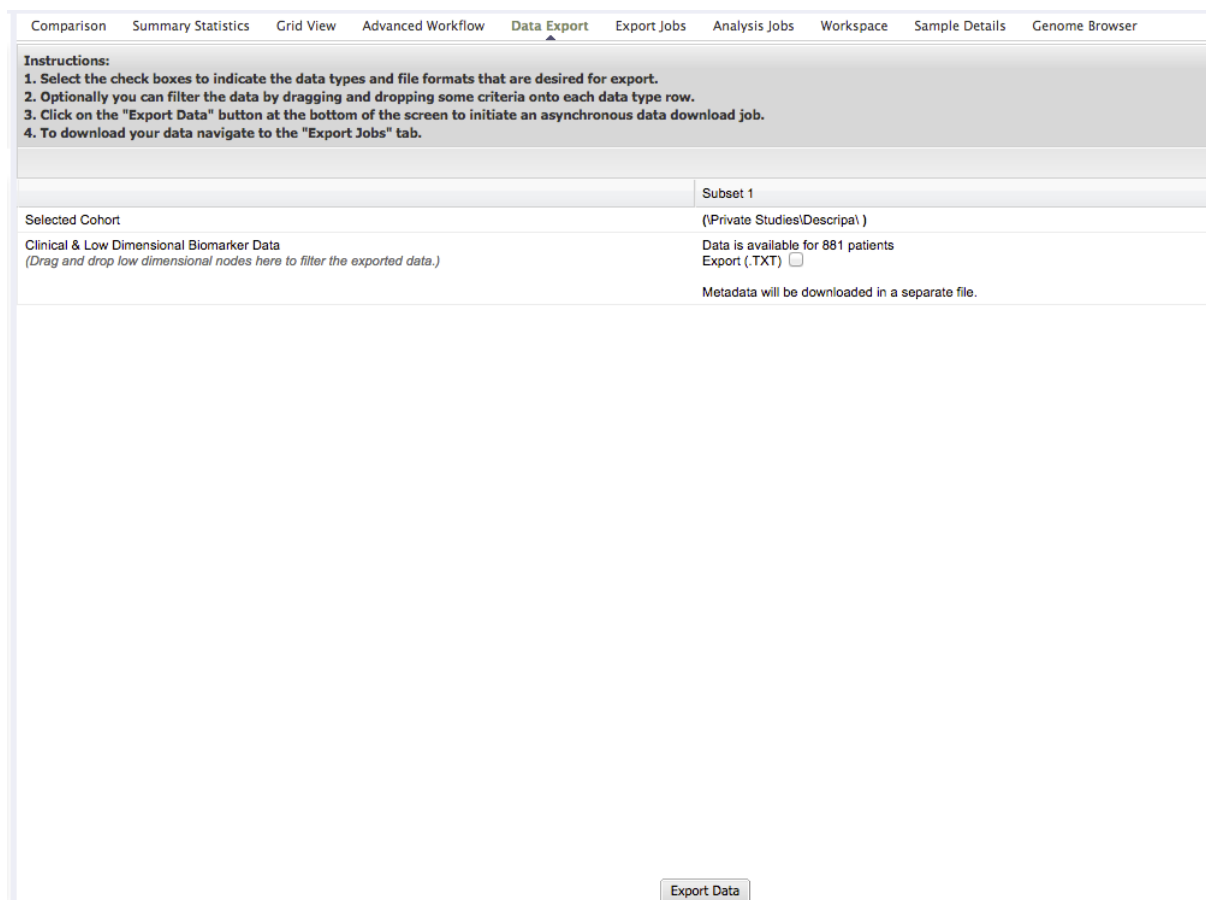


Figure 9. TranSMART Data Export form.

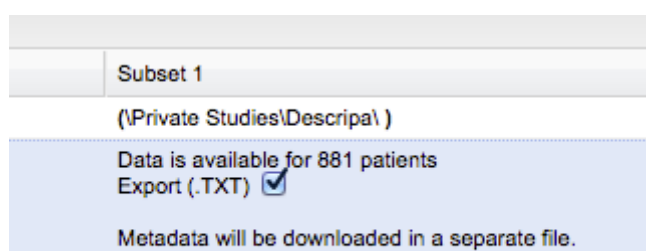



Figure 10. TranSMART Data export selections.

It is possible also to run the export job in the background in order to continue with other analyses and cohort selection while the job completes. The job could take several minutes depending on the amount of data selected.

Analyses run through the Advanced Workflow tool within Dataset Explorer use R for computation. It is possible to download raw R data for use in an external tool. After the analysis is completed, links with downloadable raw R data appear on the screen.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools	Version: v1.0 – Final	
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	19/21

Data export can for example be useful for a structured, procedural analysis that consists of several, consecutive steps. The analyses in tranSMART are oriented towards exploratory interactions with the data.

3. Framework for R plugins

Framework for R plugins gives an opportunity to extend the existing analysis types in tranSMART and to develop new analysis if needed.

Any of the advanced analyses built-in in tranSMART invoke an R-script in the back-end. In general, new analyses can be built in by integrating the R code into the existing code, and make sure input (data format) and output (data format + e.g. graphs/tables) are defined. Also a basic graphic interface needs to be designed, which can capture the input parameters that need to be set by the end-user. A technical description of the steps involved is available upon request by the tranSMART team, who supports tranSMART within EMIF, as part of the technical documentation of tranSMART.


4. Cross Trial Query

TranSMART allows performing cross trial comparisons, where several studies are merged into a single, harmonized virtual dataset. One of the prerequisites for cross trial queries is to load study data in a particular manner. We are currently reviewing the application to assess which parts of the application support cross trial data, and which parts are limited to single cohorts only.

4.1. Cross Trial Queries in TranSMART

Screenshots included below are taken from a pooled dataset consisting of three different study datasets (AddNeuroMed, Descripa, SantPau). Data were pooled outside of tranSMART before upload. The idea of cross trial queries functionality in tranSMART is to have a *virtual* pooled dataset, instead of uploading data twice.

Figure 11 represents a pooled dataset, including a variable showing the cohort where data originated. This variable can be used in the analysis to distinguish datasets, and do cohort comparisons.

 IMi - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)	Security: PU	20/21

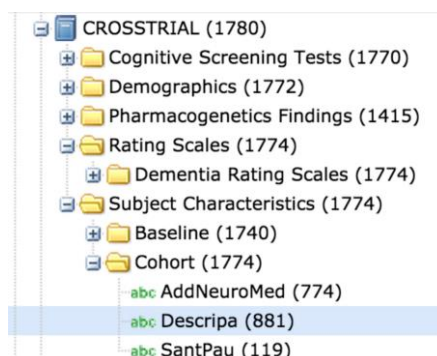


Figure 11. Three cohorts represented as a single pooled dataset.

4.1.1. Example 1: Compare two cohorts using Summary statistics

TranSMART allows users to quickly view comparisons between cohorts, via the Summary statistics. Figure 12 compares age distributions between AddNeuroMed and Descripa.

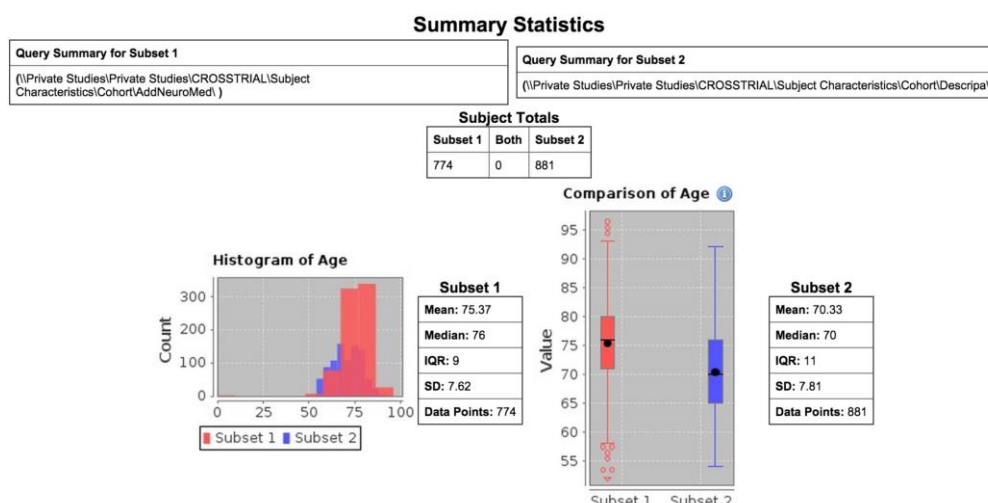



Figure 12. Summary statistics for two cohorts.

4.1.2. Example 2: Compare clinical rating scale results between two cohorts

Additional variables can be dragged from the taxonomy to compare these variables for the selected cohorts.

 IMI - 115372	D13.2 Data analysis tools for vertical projects version 1		
	WP13 Analysis, processing & visualization methods and tools		Version: v1.0 – Final
	Author(s): Janneke Schoots – van der Ploeg (The Hyve for JANSSEN), Rudi Verbeeck (JANSSEN)		Security: PU 21/21

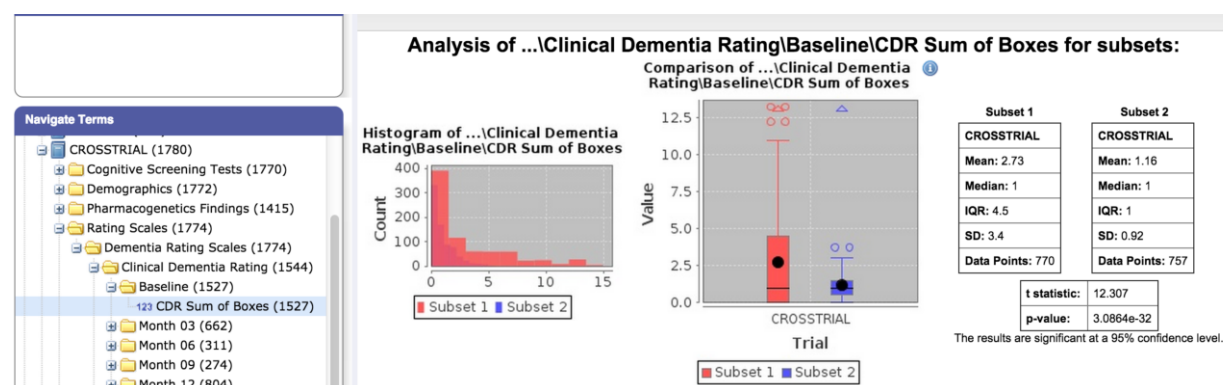


Figure 13. Cross cohort score comparison.

4.1.3. Example 3: Linegraph showing time course data per cohort

Figure 14 shows average MMSE scores over time (4 visits) for AddNeuroMed and Descripa.

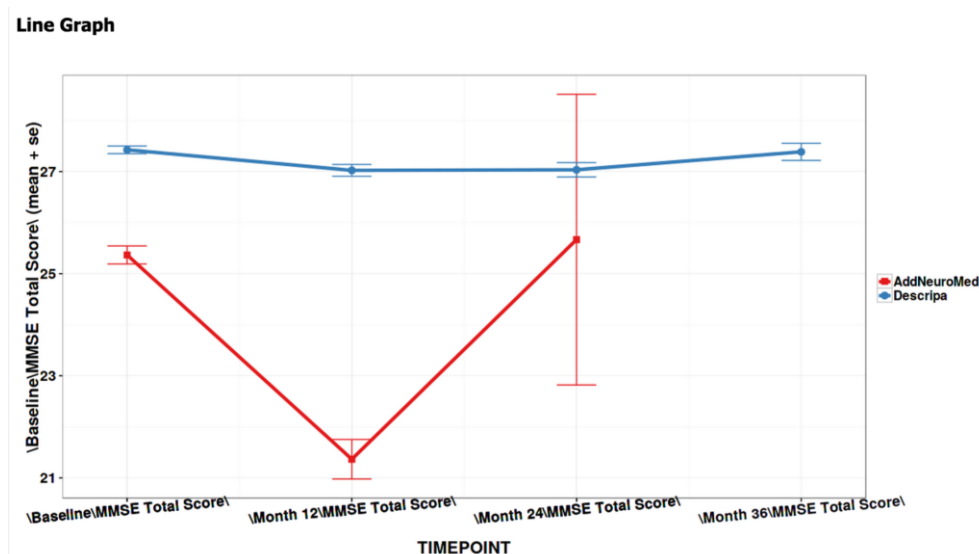


Figure 14. Cross cohort linegraph.

4.2.Data load for cross trial queries

The scripts for uploading clinical data do support loading associations between the clinical data variables and the Across Trials variables. For the across trials data to be loaded, an extra file is needed (henceforth "the across trials file"), in which an exact mapping between study variables and cross trial variables is provided. Prerequisite for loading this type of data is to run the data-loading job using a newly implemented ETL framework, which is called tranSMART-batch.