Integrative Cancer Research
INTEGRATE
Through Innovative Biomedical Infrastructures

SEVENTH FRAMEWORK PROGRAMME

ICT-2010-270253

INTEGRATE

Driving excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures

STREP
Contract Nr: 270253

D5.3 Report on the modelling framework and the predictive models for therapy response

Due date of deliverable: (01-08-2013)
Actual submission date: (MM-DD-YYYY)

Start date of Project: 01 February 2011 Duration: 36 months

Responsible WP: FORTH

Revision: <outline, **draft**, proposed, accepted>

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	X

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
George Manikis	FORTH	gmanikis@ics.forth.gr
Evaggelia Maniadi	FORTH	maniadi@ics.forth.gr
Kristof De Schepper	Custodix	kristof.deschepper@custodix.com
Jelle Vandriessche	Custodix	jelle.vandriessche@custodix.com
Njin-Zu Chen	Philips	njin-zu.chen@philips.com
Kostas Marias	FORTH	kmarias@ics.forth.gr

0.2 Documents history

Document version #	Date	Change
V0.1		Starting version, template
V0.2		Definition of ToC
V0.3	29/10/2013	First complete draft
V0.4		Integrated version (send to WP members)
V0.5		Updated version (send PCP)
V0.6		Updated version (send to project internal reviewers)
Sign off		Signed off version (for approval to PMT members)
V1.0		Approved Version to be submitted to EU

0.3 Document data

Keywords	
Editor Address data	Name: Partner: FORTH Address: Phone: Fax: E-mail:
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer
		fp7-integrate@listas.fi.upm.es

Table of Contents

0 DOCUMENT INFO 2
0.1 Author 2
0.2 Documents history 2
0.3 Document data 2
0.4 Distribution list 2
1 INTRODUCTION 4
2 INTEGRATE ANALYSIS FRAMEWORK 5
2.1 Introduction 5
2.2 Architecture & Specifications of the Developed Framework 5
2.2.1 THE CORE FUNCTIONALITY 5
2.2.2 SECURITY AND SINGLE-SIGN-ON (SSO) 7
2.2.3 WEB SERVICES FOR RETRIEVING DATA 8
2.2.4 OVERVIEW OF THE PORTLETS 9
2.2.4.1 User Authentication 9
2.2.4.2 Data Sources 9
2.2.4.3 Analytical Tools portlet for cohort selection 10
2.2.4.4 Predictive Models portlet 14
2.2.4.5 History 17
2.2.5 INTEGRATION WITH OTHER TOOLS WITHIN INTEGRATE 19
2.3 Addressing Clinical Scenarios through the INTEGRATE Analysis and Prediction Framework 20
2.3.1 USING THE INTEGRATE ANALYSIS FRAMEWORK FOR STATISTICAL ANALYSIS 21
2.3.1.1 “Assessing the variability, dependency and the distribution of certain clinical characteristics across patient population” 21
2.3.1.2 “Making comparison tests and evaluating response rate of different examined regimens to a certain patient cohort” 23
2.3.1.3 Defining if specific clinical parameters are surrogate markers for the survivability of a patients' group, involving the modeling of time to event data in survival analysis 24
2.3.1.4 Performing quality control tests to the genomic data and identifying statistically significant genomic information through unsupervised learning 25
2.3.2 USING THE INTEGRATE ANALYSIS FRAMEWORK FOR PREDICTIVE MODELLING 29
3 SUMMARY 33

1 INTRODUCTION

The main objectives of this work package (WP5) are a) to propose an approach and a methodology for developing multi-scale predictive models in breast cancer and b) to build a corresponding framework for deriving such models based on the multi-level heterogeneous data provided by clinical trials in the neoadjuvant setting. The models developed in this work package are based on realistic clinical research scenarios which have been developed based on the neoBIG research program, addressing needs from rigorously conducted breast cancer clinical trials. The developed predictive analysis framework featuring a comprehensive clinical trial data-viewer has been largely driven by the clinical scenarios for the INTEGRATE VPH use (see D.1.2) as well as a number of discussions with the bioinformaticians/clinicians involved in the project to ensure that it will address the clinical needs. It allows scientists from diverse backgrounds to employ with ease (at the push of a button) a) sophisticated statistical analysis tools that play and important role in deeply understanding and preparing the available multi-level data for further analysis, and, b) to derive predictive models (again at the push of the button) from clinical trial data.

It is important to keep in mind that the goal of WP5 is to deliver the tools and the framework for creating and validating the models. The framework developed aims to:

- Assist users in employing the statistical analysis tools implemented within the framework, addressing specific clinical questions.
- Define the framework which provides the users the tools needed to construct and validate their own predictive models within the context of BIG clinical trials.

However, it is important to clarify that ***the clinical validation of any given model is out of the scope of this project*.** Such validation will be organized at a later stage within the context of specialized trials within the participating clinical sites.

2 INTEGRATE ANALYSIS FRAMEWORK

2.1 Introduction

The INTEGRATE Analysis Framework is dedicated in providing users with a web-based access to a collaborative, multi-functional and easy-to-use environment for exploiting, analyzing and assessing the quality of large multi-level data. The main goal is to empower the clinician to analyze with ease clinic-genomic data in order to get simple statistics on selected parameters, perform survival analyses, compare regimens in selected cohort of patient, obtain genomic analysis results, and construct predictive models using homogeneous and/or heterogeneous large multi-modal data.

The major advantage of this framework is that brings all the functionality needed for biomarker selection and model testing within a single easy to use framework that does not require knowledge of any specific software environment (e.g. Matlab, R) while it offers all the functionality needed within simple menus and buttons allowing non-experts to perform statistical analysis and modeling tasks on their data. The following sections describe the architecture and the functionality of the framework while section 2.3 of this deliverable presents indicative results and guidelines for using this framework for statistical analysis and predictive modeling from patient cohort data.

2.2 Architecture & Specifications of the Developed Framework

2.2.1 The Core Functionality

The idea behind the INTEGRATE Analysis Framework is to provide users with a web-based interface that supports user authentication/authorisation, data handling, execution of the tools and models, and visualization and storage of the analysis reports. To achieve this goal, the programming aspects of the different environments and languages adopted for implementing the framework's facilities, and the connectivity process which allows the interaction between these components are kept at the back-end of the framework, hiding the complexities of the computational infrastructure.

The frond-end of the framework, hiding the complex infrastructure is based on the Liferay Portal[1]. Liferay Portal is an enterprise web framework based on Java technologies. Our decision to choose a third party open source portal mechanism as the base of the INTEGRATE Analysis Framework, rests to the fact that there is a devoted base of developers who has adopted a frequent cycle of updates, where in each cycle numerous updates such as security enhancements, optimizations and adoption of new web technologies are provided. Another key-point is that by upgrading and extending the functionalities of the Content Management System (CMS) we are able to provide a consistent user interface for all the modules of the framework ensuring a seamless user experience.

[1] Liferay (www.liferay.com)

The front-end of the framework is enhanced with JavaServer Faces (JSF), a Java technology for building component-based user interfaces for web applications. JavaServer Faces technology simplifies building user interfaces for JavaServer applications. Various Ajax-based JSF frameworks and a wide variety of components exist in several libraries (e.g. RichFaces, ICEfaces, PrimeFaces). Among these libraries, PrimeFaces[2] was chosen to be used in the INTEGRATE Analysis Framework. PrimeFaces is a lightweight open source component suite for JavaServer Faces, offering over 100 individual components (mostly visual), covering a diverse range of widgets including Ajax, input fields, buttons, data display controls, panels, overlays, menus, charts, dialogs, multimedia presentations, drag/drop and other controls.

The core functionality of both the statistical and predictive analysis is written in R language[3] using publicly available libraries from a large repository. R provides a powerful suite of tools for analysis, a highly extensible coherent system for software development, and a good connectivity with other software environments. To facilitate embedding R functionality in our java-based interface, a client/server concept using TCP/IP[4] protocol was used for the communication between the R system and the end-user allowing interaction between the framework and the execution environment. At the same time connections between multiple clients-users and the R system are established using their own data space and working directory without interfering with other connections.

For each running statistical analysis tool or predictive model, the INTEGRATE Analysis Framework supports an engine[5] to create dynamically analysis reports by enabling integration of R code and Latex documentation[6]. On-the-fly reporting is therefore generated by combining the programming source code and the corresponding documentation into a single file. At last, the framework provides an internal database where a full analysis record of an executed analysis is stored, including metadata information such as timestamp information, tool/model authorship, type of the analysis, the examined data, any memory constraints, the analysis progress (complete or pending), the dynamically generated reports in both .pdf and .html format, and etc.

[2] http://primefaces.org/
[3] The R project for Statistical Computing (www.r-project.org)
[4] Rserve, a binary R server (www.rforge.net/Rserve)
[5] The Sweave tool (www.statistik.lmu.de/~leisch/Sweave)
[6] Latex, a document preparation system (www.latex-project.org)

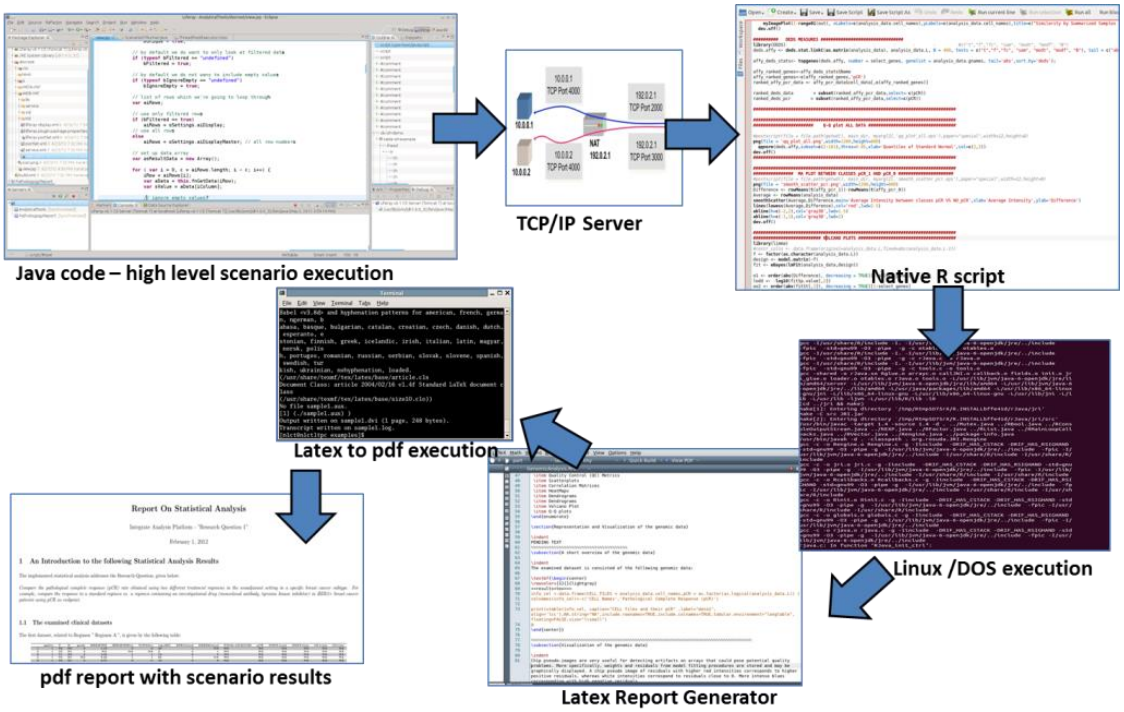

Figure 1 The back-end functionality of the Analysis Framework

Besides the internal infrastructure of the Analysis Framework, other services were implemented that contribute in securing the framework (only authenticated and authorised users can have access), as well as allowing the analysis framework to achieve communication with external data central repositories for retrieving data for analysis. The following two chapters give a brief overview of the used functionality.

2.2.2 Security and Single-Sign-On (SSO)

The INTEGRATE Analysis Framework relies on the overall INTEGRATE security framework for enabling authentication and (basic) authorisation. For enabling the authentication, the Liferay standard authentication modules are extended and connected to the central Identity Provider (IdP), which is part of the INTEGRATE identity framework. As seen in D.2.6, this IdP provides an implementation of the SSO browser based SAML profile.

If a user tries to access one or more protected resource(s) on the INTEGRATE Analysis Framework (1), he is redirected to this IdP (2) who will issue a security token after the user has authenticated him/herself to the system (3). This security token will then be validated by the local authentication module (4) of the framework, and the obtained validation result is used to make an access decision for that specific user (5).

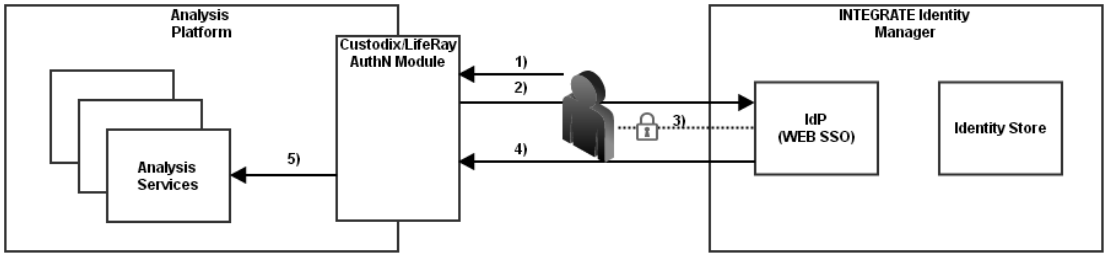

Figure 2 SSO Authentication in INTEGRATE Analysis Framework

2.2.3 Web Services for retrieving Data

The Semantic Interoperability Layer is divided in 2 main components; the Common Data Model (CDM) and the Core Dataset. CDM acts as the data model of the framework and the Core Dataset is the medical vocabulary of the framework. These two components form the Common Information Model (CIM). The main goal of the CIM is to provide homogeneous access to the different data sources. Around these components a set of services have been developed to facilitate the process of loading and retrieving data from a common data infrastructure for clinical data.

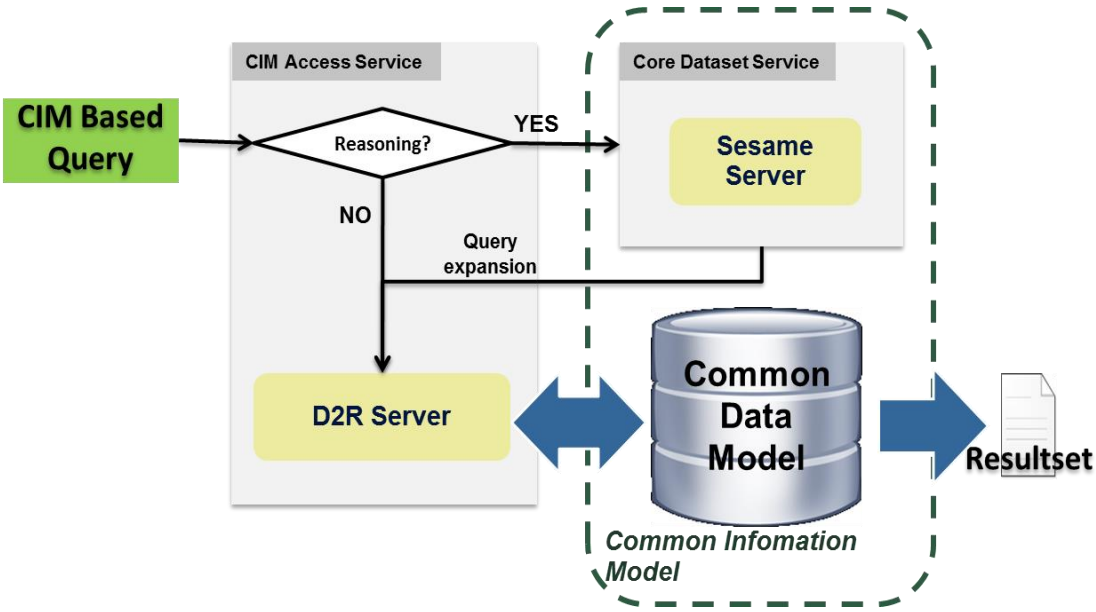

Figure 3 Semantic Layer retrieving data process[7]

In this context, the process of retrieving data is comprised by two services (CIM access service and Core Dataset service) and the data warehouse, as it is presented in Figure 3. In summary, the CIM Access Service receives a SPARQL query based on

[7] Deliverable 2.6: System Architecture Refinement, Security Framework and Implementation Status

CDM. This query is expanded with data from the core dataset service if needed and executed against the CDM. Finally, the results are returned in RDF format[8].

2.2.4 Overview of the Portlets

In the following chapters we give an overview of the implemented portlets. In general, each portlet plays a specific role in the INTEGRATE Analysis Framework, starting from the "user authentication" portlet for getting access to the framework, the "data sources" for interacting and retrieving the analysis data from the CDM, the "analytical tools" for performing the statistical analysis, the "predictive models" for doing the predictive analysis, and finally the "history" portlet for accessing the internal database of the framework where a full analysis record of an executed analysis is stored.

2.2.4.1 User Authentication

The user authentication process is a prerequisite step before a user accesses the INTEGRATE Analysis Framework. This can be achieved by enabling Single-Sign-On (SSO) techniques as described in 2.2.2 and depicted in Figure 4.

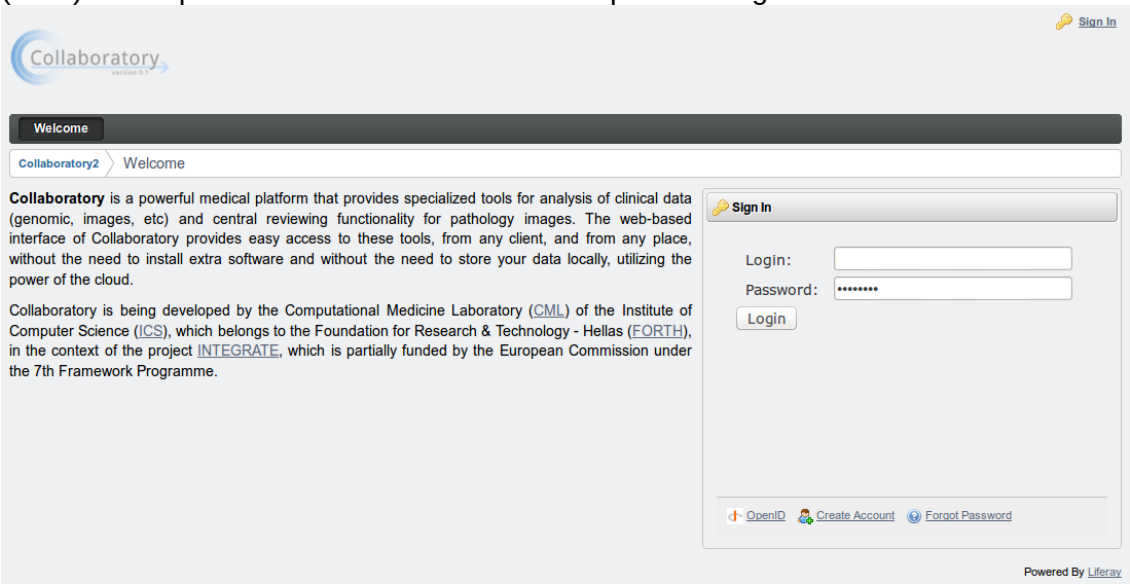

Figure 4 Main layout of the framework and SSO connectivity

2.2.4.2 Data Sources

As described in 2.2.3, the INTEGRATE architecture concerning the data storage, handling and sharing, relies on the fact that all data are stored in a central data repository named as CDM. To achieve interoperability between the INTEGRATE Analysis Framework and the Common Data Model, a web service is developed. As seen in Figure 5, the framework incorporates functionality for accessing the data stored in the CDM by the following ways:

- Non- scheduled process for retrieving directly the chosen available datasets

[8] Deliverable 3.5: Initial prototype of the semantic interoperability layer

- Scheduled process via a timetable where the user gains access to the chosen data at a specific date

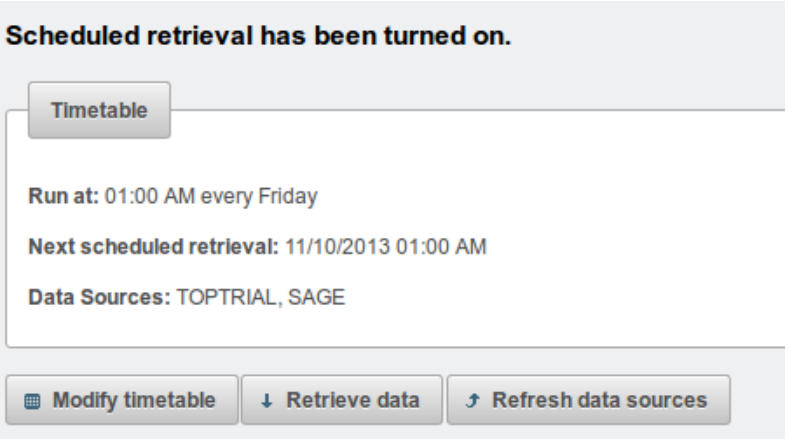

Figure 5 Schedule data retrieval via web services

During data retrieval, the necessary queries are built by the framework and sent to the CDM over the web service. Then the queries are executed and the semantic interoperable information is returned to the framework by the semantic interoperability layer. Once the data is retrieved, the user can execute the provided tools and models.

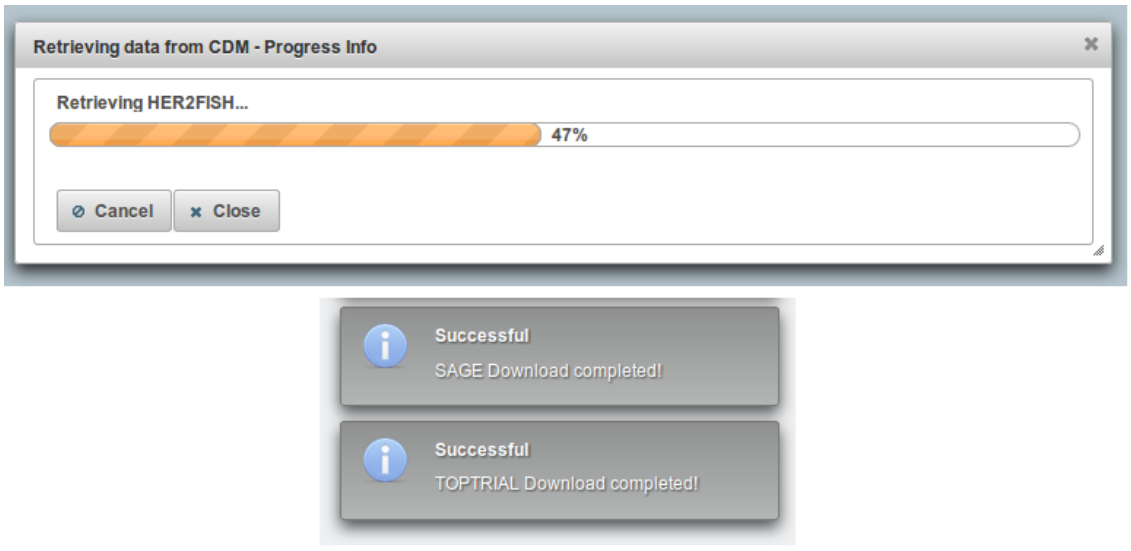

Figure 6 Getting access to the selected data

2.2.4.3 Analytical Tools portlet for cohort selection

The “analytical tools” portlet is one of the major portlets of the INTEGRATE Analysis Framework. This portlet assists users in following a pipeline process in order to perform statistical analysis in a pre-selected cohort of the retrieved dataset. Precisely, using widgets from the PrimeFaces library, the user selects a dataset that has been previously retrieved from the CDM (see Figure 7) and then proceeds to the next step where a cohort from the dataset is selected for analysis.

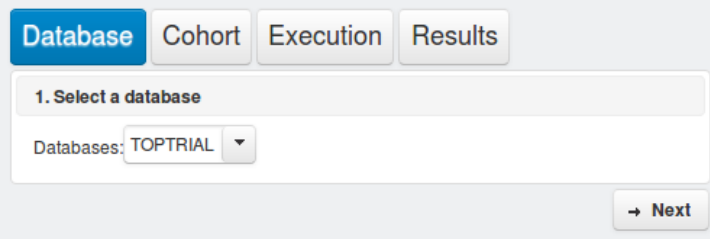

Figure 7 Selecting a dataset for analysis

In the cohort selection process, at the time the dataset is selected for analysis, the framework uses the predefined data model that identifies the specific dataset, retrieves automatically the name of each variable and displays the entire information in a widget-based table. From this table, the user selects the set of variables that will be used in the analysis (i.e. T, pCR, etc.), filters the selected variables by selecting specific ranges of values (i.e. tumor grade T1 and T2 from variable T), and finally can view the constructed cohort in a table format (see Figure 8).

At the time the cohort is been generated, the user proceeds to the next tab of the widget where the entire statistical analysis tools are displayed. All the available statistical analysis tools are presented in a user friendly manner where the user just puts a tick to the analysis that wants to be executed. The INTEGRATE Analysis Framework is then coupled with security controls that based on the type of each analysis (i.e. univariate descriptive statistics, genomic analysis, etc.) the number of variables need to be chosen for the proper execution of an analysis (i.e. bivariate descriptive statistics require two input variables), the grouping of some variables in specific lists of the widget (i.e. in survival analysis optional categorical variables can be used to produce different survival curves), and the eligible analysis criteria in general (i.e. survival analysis needs as mandatory variables the time of the event and the event itself), can protect even the non-expert user from selecting non-eligible variables for an analysis.

Conclusively, the widget-based framework assists users in selecting multiple statistical analysis scenarios (i.e. survival analysis, univariate descriptive statistics, etc.) and several executions of the same scenario using more than a single variable/group of variables (i.e. univariate descriptive statistics using “T” and “TOPOIHC”) in a single step. The overall functionality of this tab is depicted in Figure 9.

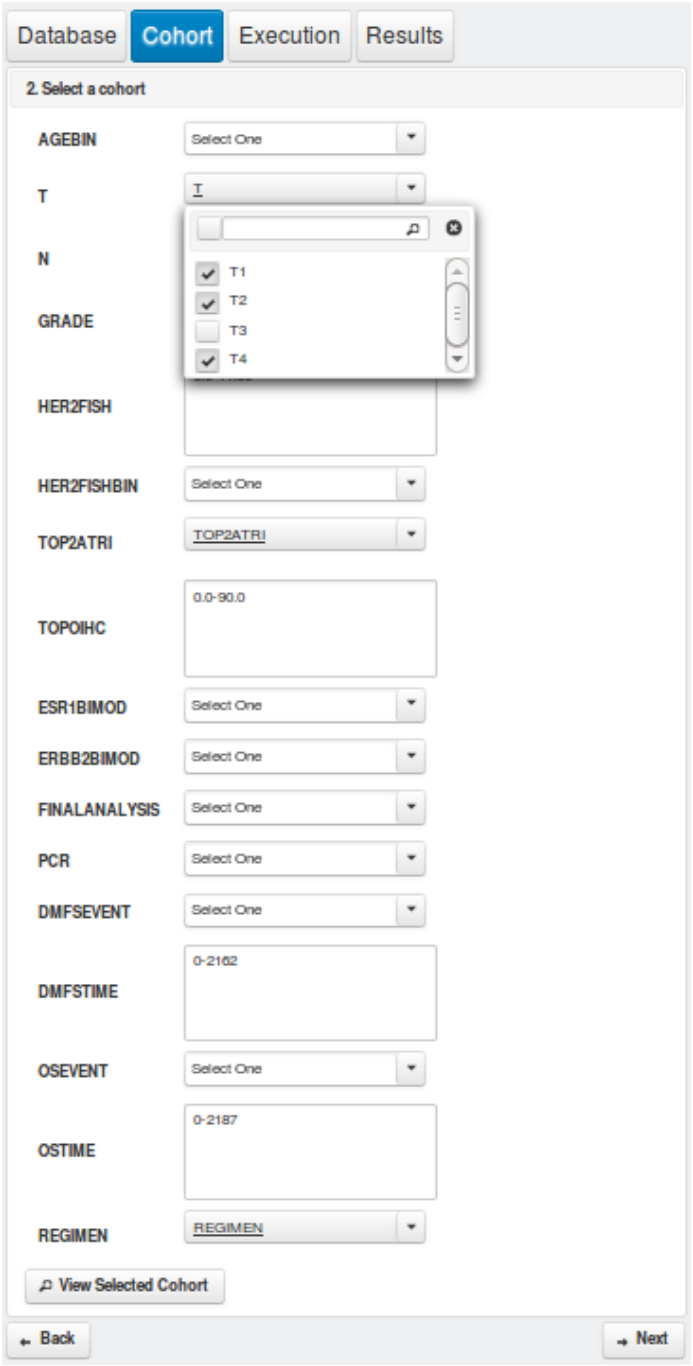

Figure 8 Selecting the examined variables for analysis

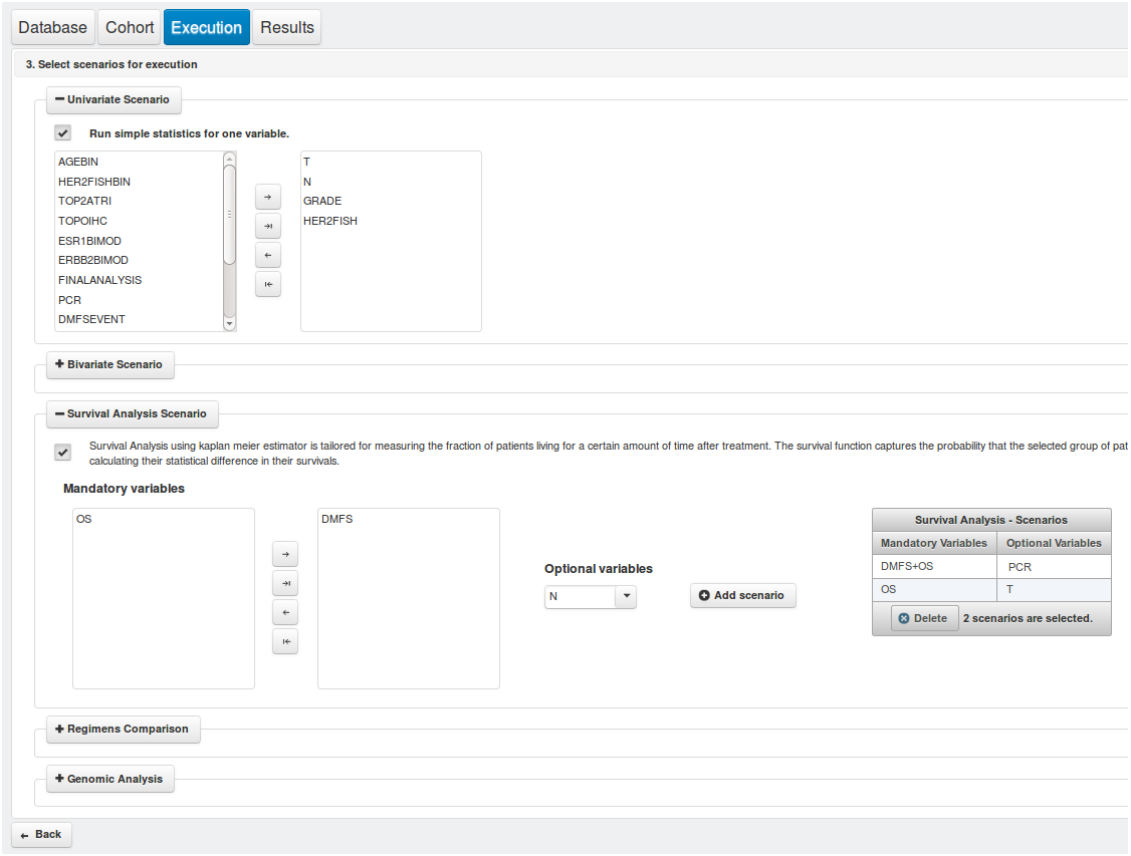

Figure 9 The statistical analysis main layout

Finally, the layout communicates with the back-end functionality and the required software, and the overall analysis workflow is presented in a diagram format. Every component of the diagram is functional and when clicked presents the information that corresponds to it. The overall diagram, based on the colour of each component, is separated into the following stages:

- Cohort selection process (yellow components in Figure 10)
- Type of analysis (blue components)
- The analysis on a specific variable/group of variables (green components)

When the component that corresponds to the cohort selection process is pressed, a pop-up window appears and the generated cohort (see Figure 8) is displayed in a table format. If the user wants to see the statistical results of a specific analysis, then each green component, related to a specific analysis, is connected to the on-the-fly generated report and the full analysis report is displayed in a pdf format.

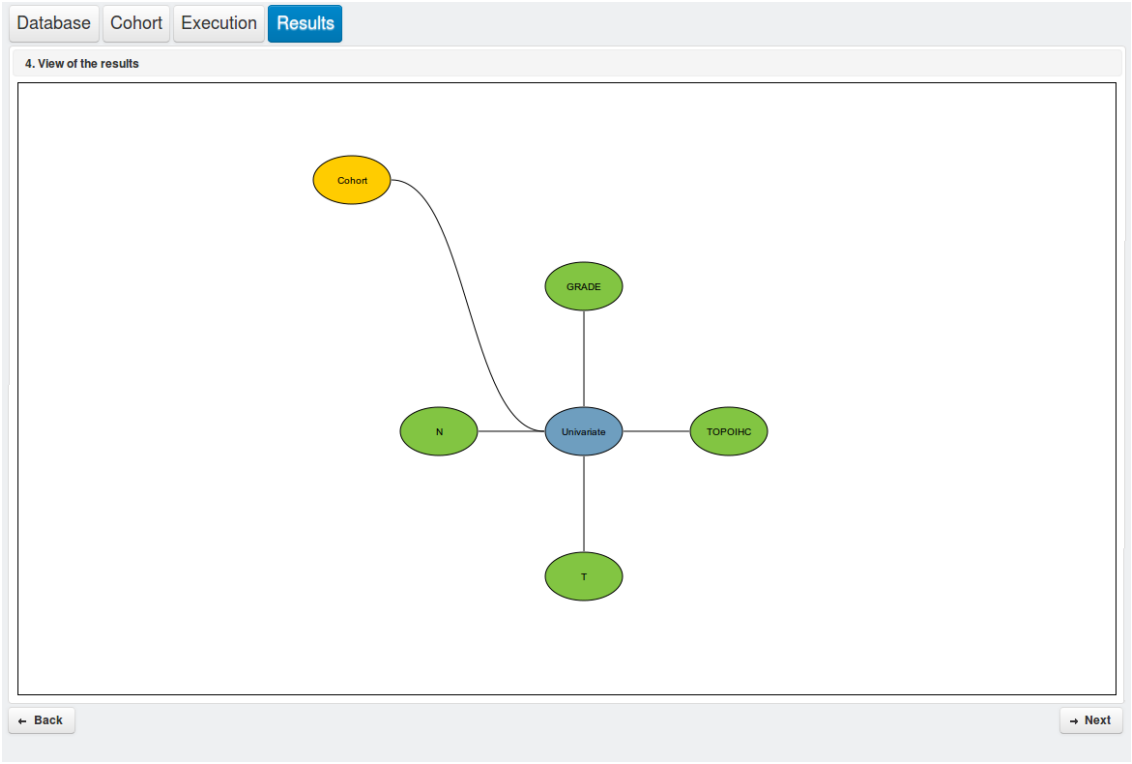

Figure 10 Displaying the entire workflow of the analysis

2.2.4.4 Predictive Models portlet

Architecturally, the predictive analysis portlet follows almost the same workflow as in the statistical analysis portlet. At first, the user selects the dataset to be retrieved by the Common Data Model and then builds the cohort for analysis which can be either homogeneous (i.e. composed of gene expression data) or heterogeneous (i.e. integrated clinical and gene expression data). Once the cohort is generated, the user proceeds to the main layout of the predictive analysis where both the model for homogeneous (Basic Predictive Model) and heterogeneous data (Advanced Predictive Model) is displayed (see Figure 11).

The predictive analysis framework implemented within the INTEGRATE Analysis Framework allows users to build a predictive model (training process) using a cohort with known clinical outcome (i.e. gene expression data and known pathological complete response for each patient), to predict the clinical outcome of a new cohort based on an already trained model, or to perform a complete predictive analysis study where at first a subpopulation of the selected cohort is used to train the model and then the rest subset is used for assessing the predictive accuracy of the model. All options in predictive analysis rely on the mechanisms that will be presented in 2.3.2.

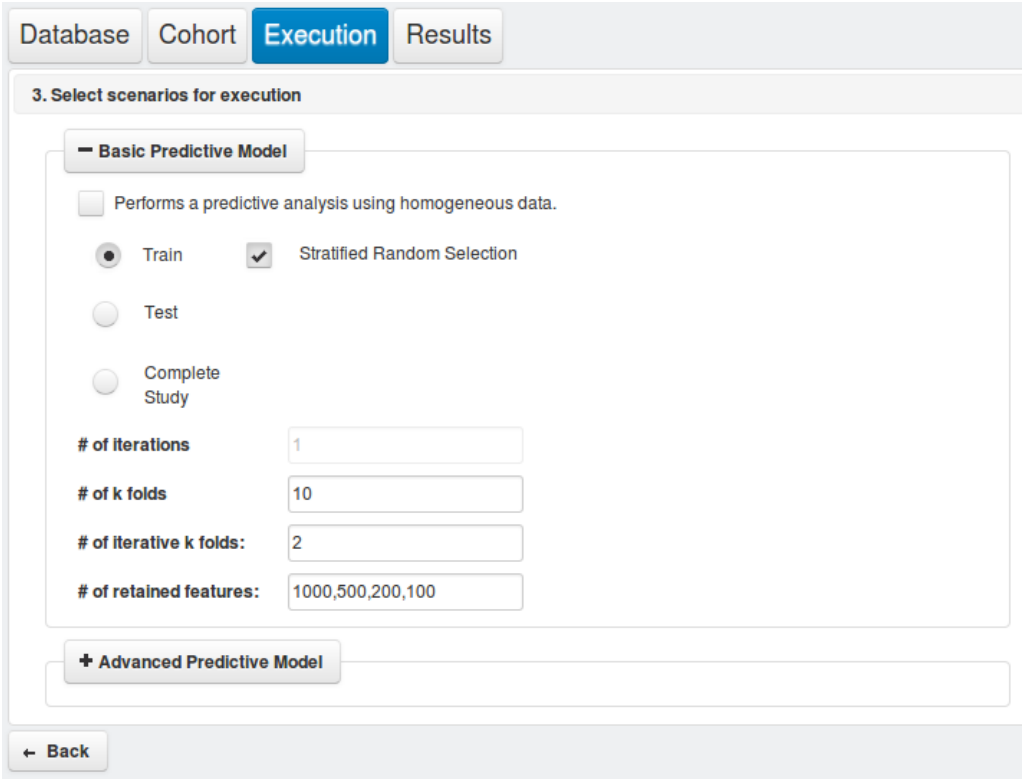

Figure 11 The predictive analysis main layout

In case the user wants to build a predictive model based on a cohort with known clinical outcome (i.e. pathological complete response), then the "Train" option is checked. Additionally, the user can select to perform a stratified random selection to the generated cohort which means that equally distributed subpopulation from the entire cohort is randomly selected to build the predictive model. The percentage of the cohort subpopulation is given by a pop-up window as depicted in Figure 12.

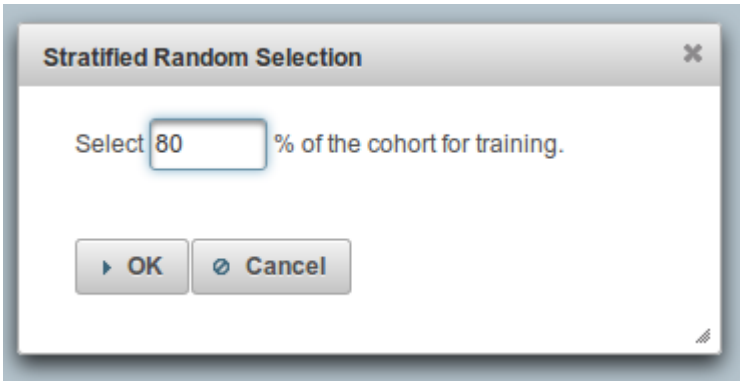

Figure 12 Training process using randomly a percentage of the entire generated cohort

When the training process is ended, the internal database of the framework keeps a full record of the analysis, containing the cohort used for building the model, the analysis reports, metadata information, etc. This procedure is part of the security controls adopted in the framework in order to ensure that every analysis is running and completed in a proper way.

In case a model has been already built and the user wants to predict the clinical outcome of a new cohort, the “Test” option is checked. Then, a pop-up window containing all the available predictive models that were built within the framework assists user in selecting the preferable to be applied to the new cohort (see Figure 13). The pop-up window contains information such as timestamp, the data used (i.e. gene expression), the clinical outcome used as the predictive outcome (i.e. survival status), if the cohort is a subpopulation of the retrieved dataset (i.e. cohort was filtered by a binary value “Age”, meaning patient less than 50 years old), and if stratified random selection was followed during the training process.

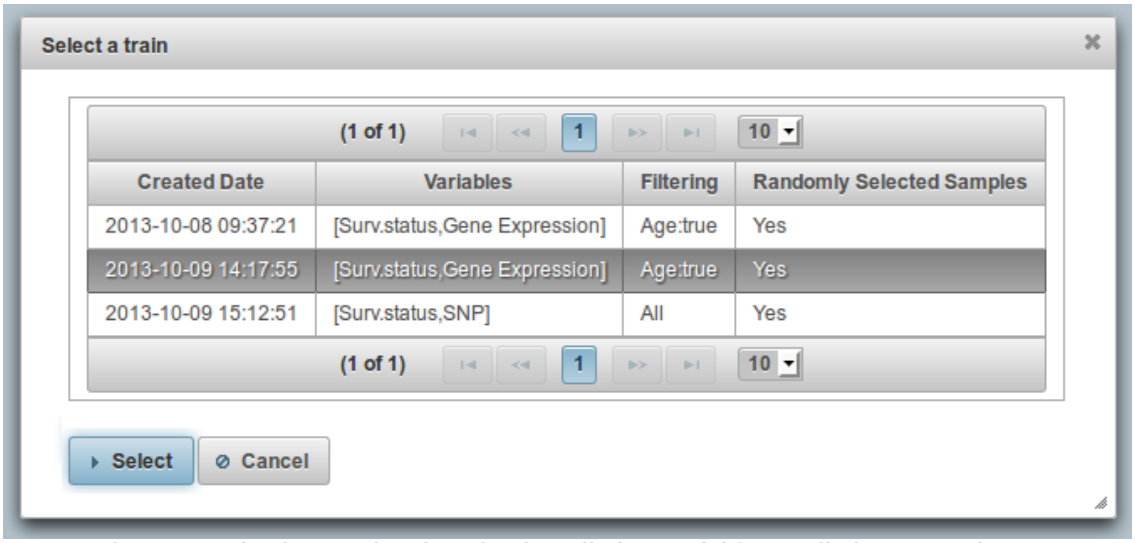

Figure 13 Selecting an already trained predictive model for predicting new cohorts

In case the user selects an already trained predictive model where gene expression data is used for building the model and performs testing of a new cohort that is not consisted of the same type of data (i.e. clinical data or single-nucleotide-polymorphism data), then a warning is displayed and the user needs to re-select the appropriate data in order to continue with the analysis.

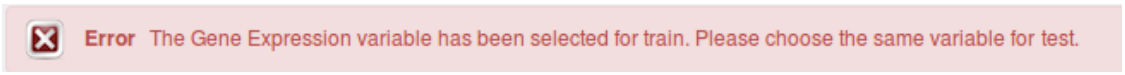

Figure 14 Displaying error while selecting a new cohort for testing

Furthermore, if the selected cohort for testing is composed of cases-patients already used during the training process (build of the predictive model), then a new pop-up window warns the user about the percentage of the cohort previously used for training. If the user wants to exclude this subpopulation from the cohort (a common technique in pattern recognition in order to avoid bias in the model), the “yes” button is clicked and the new cohort contains only “unseen”, new cases. This security control is depicted in the following figure.

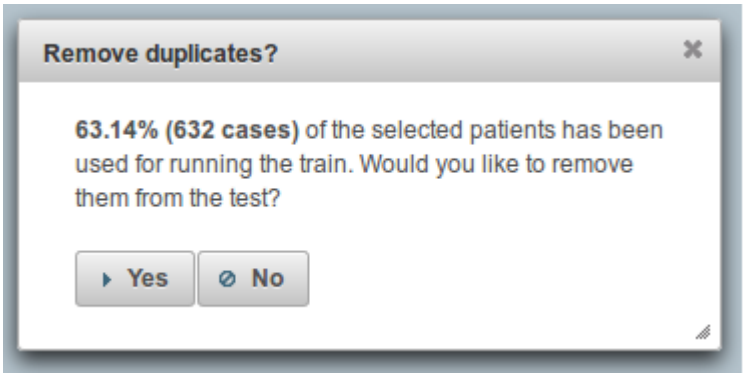

Figure 15 Excluding duplicates before predicting the outcome of a new model

2.2.4.5 History

The INTEGRATE Analysis Framework is equipped with an internal database that stores all the metadata information for every executed analysis. In other words, all users have a private space were all analysis in included, stored, and can be used for further analysis (i.e. an already built model that can be used to predict the clinical response of a new trial). Each analysis record contains the following information:

- Timestamp information (i.e. started date and time of execution).
- Tool/Model authorship
- Type of the analysis
- Selected variables
- Filtering information
- Execution time (in seconds)
- Progress of the analysis execution ("in progress" or "completed")

This metadata information is presented in a tabular format, as depicted in Figure 16. The user can navigate through the metadata information and select an analysis. For that analysis the user can a) view the report in either pdf or html format that is dynamically generated during the execution of the selected analysis, and b) view the selected cohort in a tabular format that was/is used for the execution of an analysis.

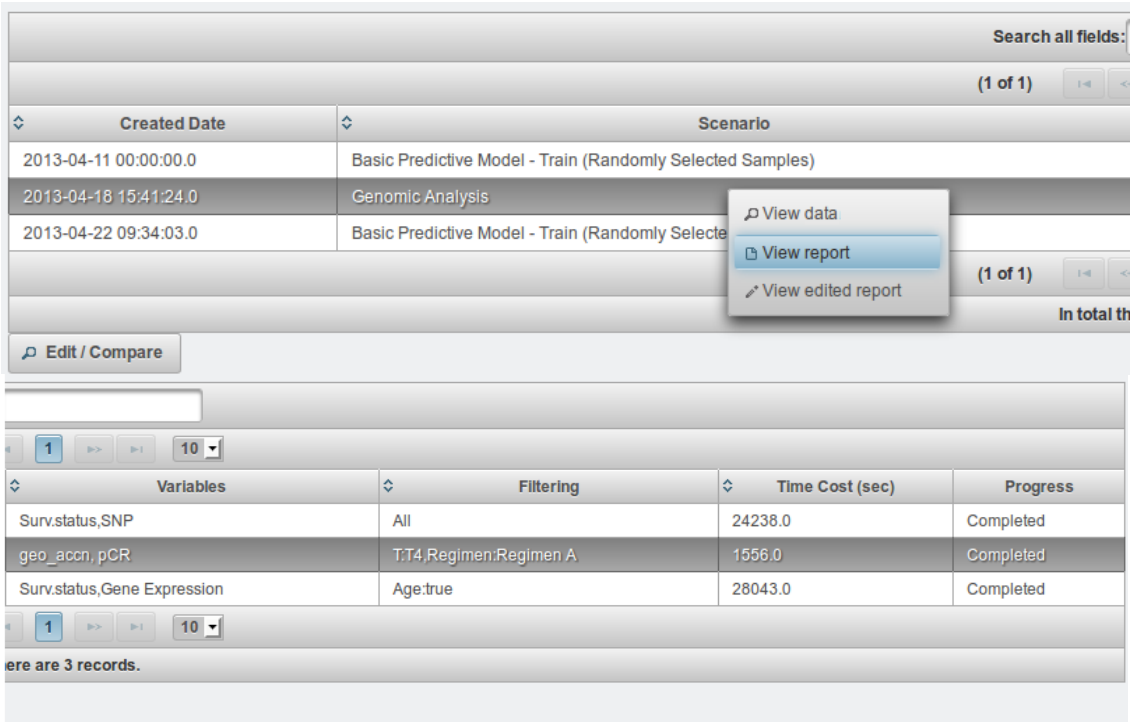

Figure 16 The main layout of the history portlet splited into two subfigures

An optional functionality is also provided by the INTEGRATE Analysis Framework where the user can compare the results from different executed models by vertically aligning the html reports in the browser. Additionally, the user can edit an html report, using a basic editing toolbar, and save the changes back to the server. A link to the original and edited pdf and html report is also provided. This is presented in Figure 17.

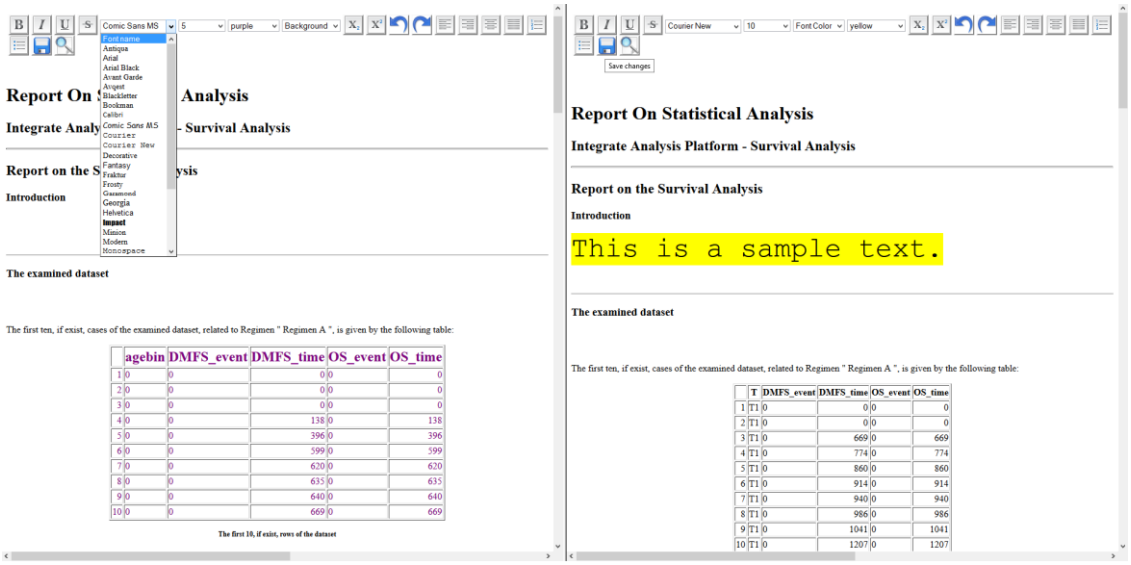

Figure 17 Compare and edit the html reports of several analyses

2.2.5 Integration with other tools within INTEGRATE

The interactive cohort selection tool is designed by our partners from Philips and is a part of the INTEGRATE where users get access to the CDM, and define cohorts on the fly, by using SNAQL scripts. These scripts can be very complex, allowing the user to find highly specific patient cohorts in the Core Dataset.

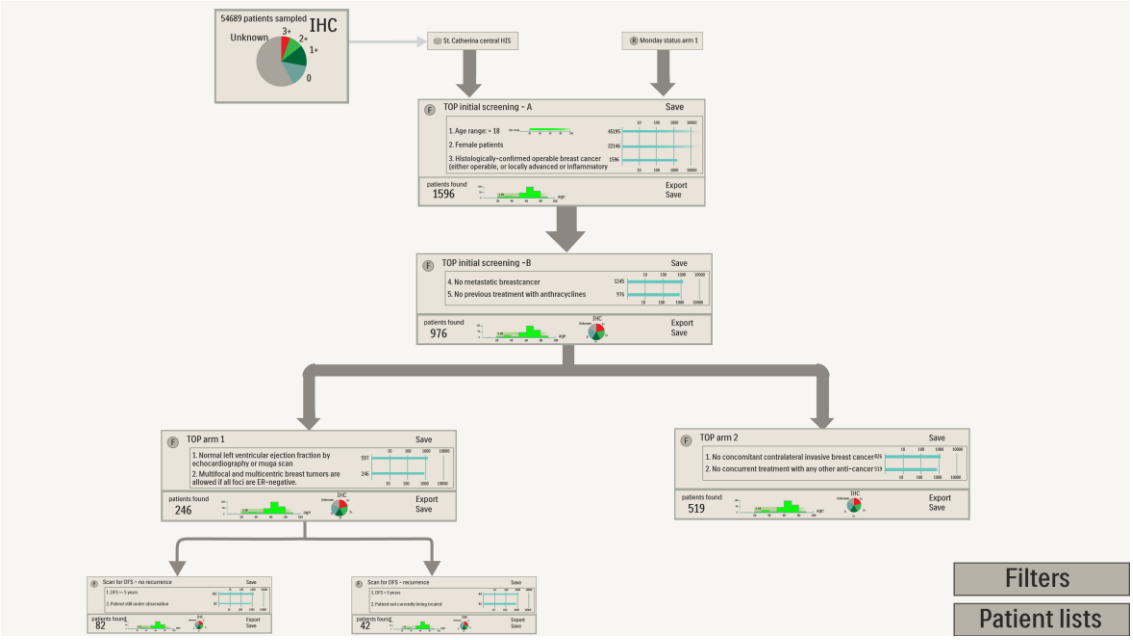

Figure 18 Defining on-the-fly cohorts using the Cohort Selection

A joint effort between FORTH and Philips partners has been paid to achieve integration between the Analysis Framework and the Cohort Selection. Using this integration, the user can select a specific group of patients within the Cohort Selection and at the same time perform the analysis to this group. This is accomplished via web services containing information about the selected cohort, the selected statistical analysis (i.e. apply descriptive statistics to tumor grading size of the selected population) and the analysis results (figures, tables, etc.). Pseudo-coding language has been generated, relating each type of the available by the framework analysis with a unique term, and incorporated into JavaScript Object Notations for parsing data structures and associative arrays of information to both frameworks. This allows the user to have more confidence in these results, and catch errors in the SNAQL scripts early on. As the confidence in the filtered cohorts grows, more elaborate analyses can be applied.

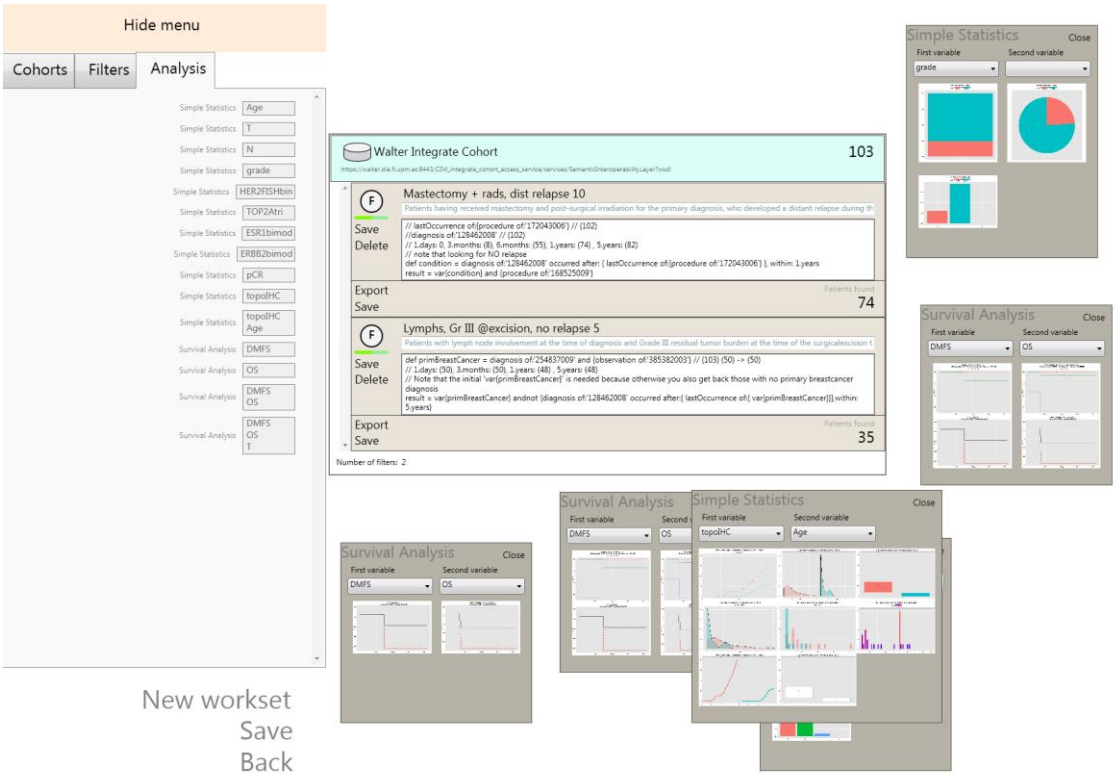

Figure 19 Indicative analysis results when using a cohort from the Cohort Analysis

2.3 Addressing Clinical Scenarios through the INTEGRATE Analysis and Prediction Framework

This section aims to shed light on the use of the analysis and modelling framework in order to derive statistics and predictive models from patient cohort data. The definition of the relevant user needs as described in D.1.2 gave rise to several research questions related to the analysis of large multi-modal data sets. These research questions or clinical scenarios alternatively, are mainly divided into two broad categories; the scenarios being covered by the statistical analysis and the predictive modelling respectively. The clinical scenarios being addressed from the statistical analysis field can be mainly grouped and focused in:

- assessing the variability, dependency and the distribution of certain clinical characteristics across patient population
- making comparison tests and evaluating response rate of different examined regimens to a certain patient cohort
- defining if specific clinical parameters are surrogate markers for the survivability of a patients' group, involving the modeling of time to event data in survival analysis
- estimating the association degree between relevant patterns, extracted from the pathology or radiology imaging data, and the clinical response of a patients' group
- performing quality control tests to the genomic data and identifying statistically significant genomic information through unsupervised learning

From the predictive modelling point of view, the implemented framework within the INTEGRATE Analysis Framework deals with the identification and assessment of gene expression signatures in predicting a specific clinical outcome (i.e. the tumour response to a specific drug used across multiple breast cancer neoadjuvant trials) and extend this predictability by assisting users in constructing predictive models using large heterogeneous data from the pool of clinical, genomic and imaging data. These can be summarized into two categories:

- an assisted predictive modeling framework when homogeneous data (i.e. gene expression) is used for building, running and evaluating the model
- a heterogeneous integration modeling framework where heterogeneous data are fused for the development of powerful multi-scale models for predicting drug response, and assessing candidate biomarkers.

2.3.1 Using the Integrate Analysis Framework for Statistical Analysis

This section covers, in a more technical manner, all the statistical analysis tools that are implemented within the INTEGRATE Analysis Framework in order to address the aforementioned, related to statistics, research questions. The aim of this chapter does not focus on the representation of the methodology and the mathematical background of each statistical tool but to highlight the statistical framework available to the user for doing a complete analysis, depending on the type and the number of variables selected for a specific cohort. As it will be discussed in the following chapters, the framework can automatically identify the type and characteristics of the selected cohort, and apply a suitable tool for the analysis. The available statistical analysis tools, as grouped in 2.3, are presented in the following topics.

2.3.1.1 Assessing the variability, dependency and the distribution of certain clinical characteristics across patient population

This clinical scenario is highly related to a broad category of statistical analysis, named as descriptive statistics, which provides simple summaries about the selected cohort in both tabular and graphical way. When the analysis involves the description of the distribution of a single variable, the framework applies univariate analysis and depending on the type of variable which is classified as numerical or categorical, this includes:

- Univariate analysis in numerical data:
  - Variable's central tendency (i.e. mean, median, etc.)
  - Dispersion (i.e. range and quantiles of the selected cohort)
  - Measures of spread (i.e. variance, standard deviation)
  - Histogram and density plot
  - Quantile-Quantile plot
  - Boxplot
- Univariate analysis in categorical data:
  - Frequency table

- Barplot
- Piechart

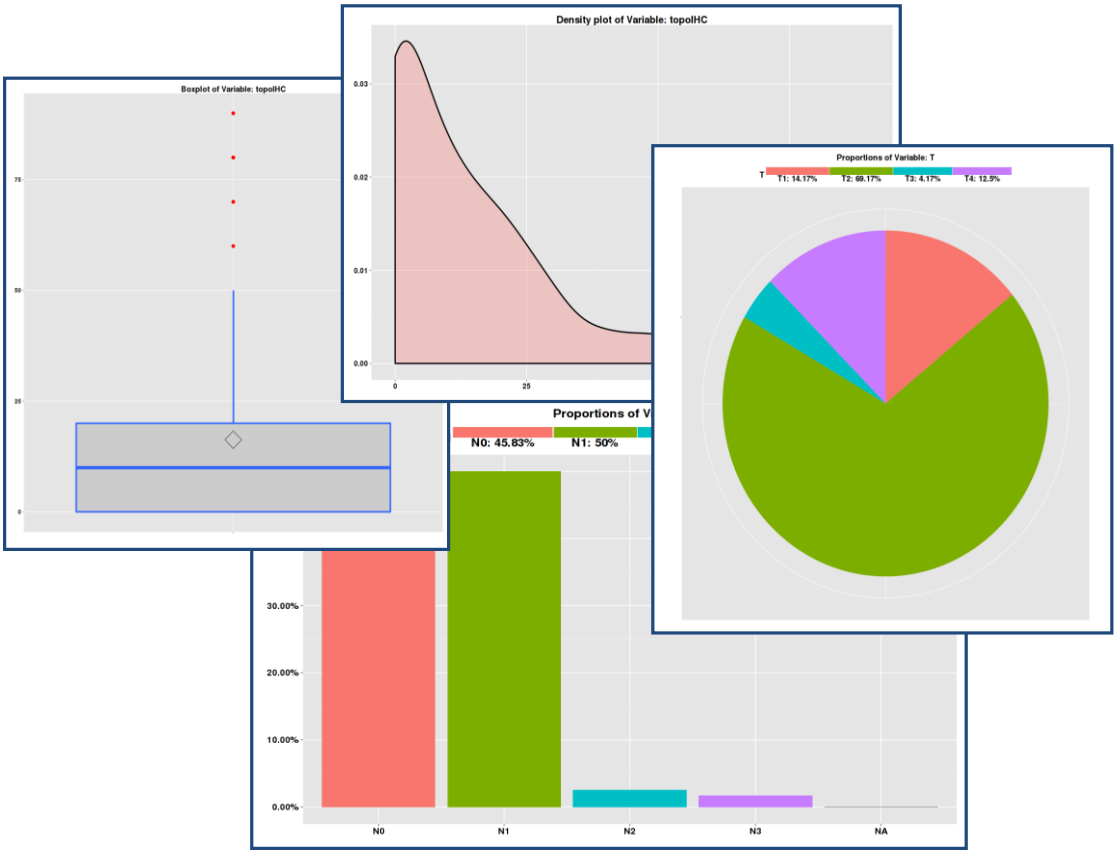

Figure 20 Indicative results using descriptive statistics in a single variable

In case the selected cohort consists of more than a single variable (a cohort composed of a pair of variables), the framework applies bivariate analysis, and depending on the combination of the variables' type (i.e. a numerical and a categorical variable, two categorical, etc.), descriptive statistics describe the relationship between the pair of variables including cross-tabulations and contingency tables, descriptions of conditional distributions, graphical representations via scatterplots, histograms, piecharts, mosaic plots and etc., and more advanced statistics like Chi-Square, Fisher, and Relative Risk tests. Indicative graphical results from the INTEGRATE Analysis Framework are depicted in the following figure.

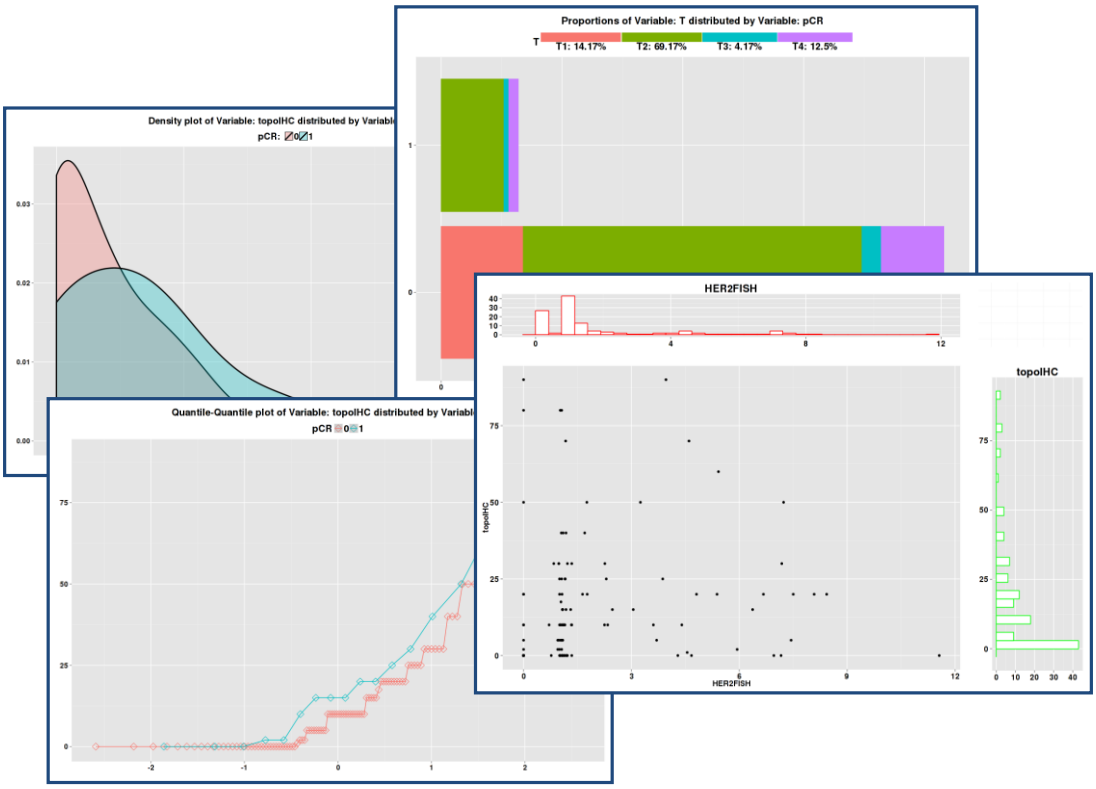

Figure 21 Indicative results using descriptive statistics in a pair of variables

Summing up, descriptive statistics within the INTEGRATE Analysis Framework can be a first-step analysis in order the user to identify any variability, dependency and the distribution of certain clinical characteristics across patient population and to have in general a more clear view about a specific cohort.

2.3.1.2 Making comparison tests and evaluating response rate of different examined regimens to a certain patient cohort

This scenario encapsulates a very first decision support system in which the response of a breast cancer subpopulation to an investigational regimen is evaluated compared to the response from a standard regimen. From the technical aspect, clinical parameters such as the pathological complete response (pCR) rate are used, and Odds Ratio though Forest Plots[9] are employed to measure the investigational regimen versus standard regimen effect.

[9] Larry V. Hedges, Ingram Olkin (1985). Statistical Methods for Meta-Analysis. Academic Press. ISBN 0-12-336380-2

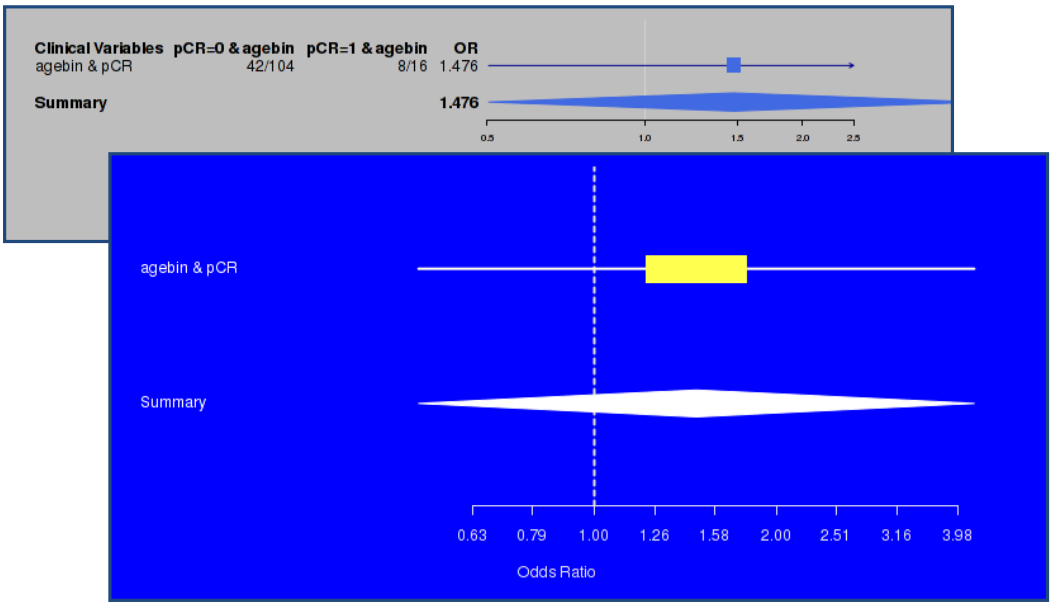

Figure 22 Indicative results presenting a number of means and their confidence intervals

2.3.1.3 Defining if specific clinical parameters are surrogate markers for the survivability of a patients' group, involving the modeling of time to event data in survival analysis

A widely used topic in statistics with a high clinical relevance is called the survival analysis. In general, survival analysis involves the modeling of time to event data, otherwise the probability of surviving at least to each time point. The INTEGRATE Analysis Framework has adopted and applied this statistical analysis into clinical studies in a sense of assessing an intervention by measuring the number of subjects survived or saved after that intervention over a period of time.

Kaplan-Meier[10] is a non-parametric method and one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment via the survival curve. This type of curve is enhanced by lower and upper confidence bounds (95% confidence bounds) in order to indicate the reliability of the estimated curve. In other words, these bounds express that 95% of the observed confidence bounds will hold the true value of the survival curve. An indicative example is given in case of studying the survivability of a selected cohort in time, taking into account any withdraws of the patients before the final outcome is observed (i.e. a survival analysis curve of a group of patients based on their pathological complete response).

The Kaplan-Meier survival analysis can be extended when the selected cohort is divided into subgroups (i.e. a cohort classified based on tumor grade, a measure of the degree of differentiation of the tumor). In this case, Kaplan-Meier curves and Chi-square tests are used to assess differences in survival between groups of subjects.

[10] Altman DG. London (UK): Chapman and Hall; 1992. Analysis of Survival times. In: Practical statistics for Medical research; pp. 365–93

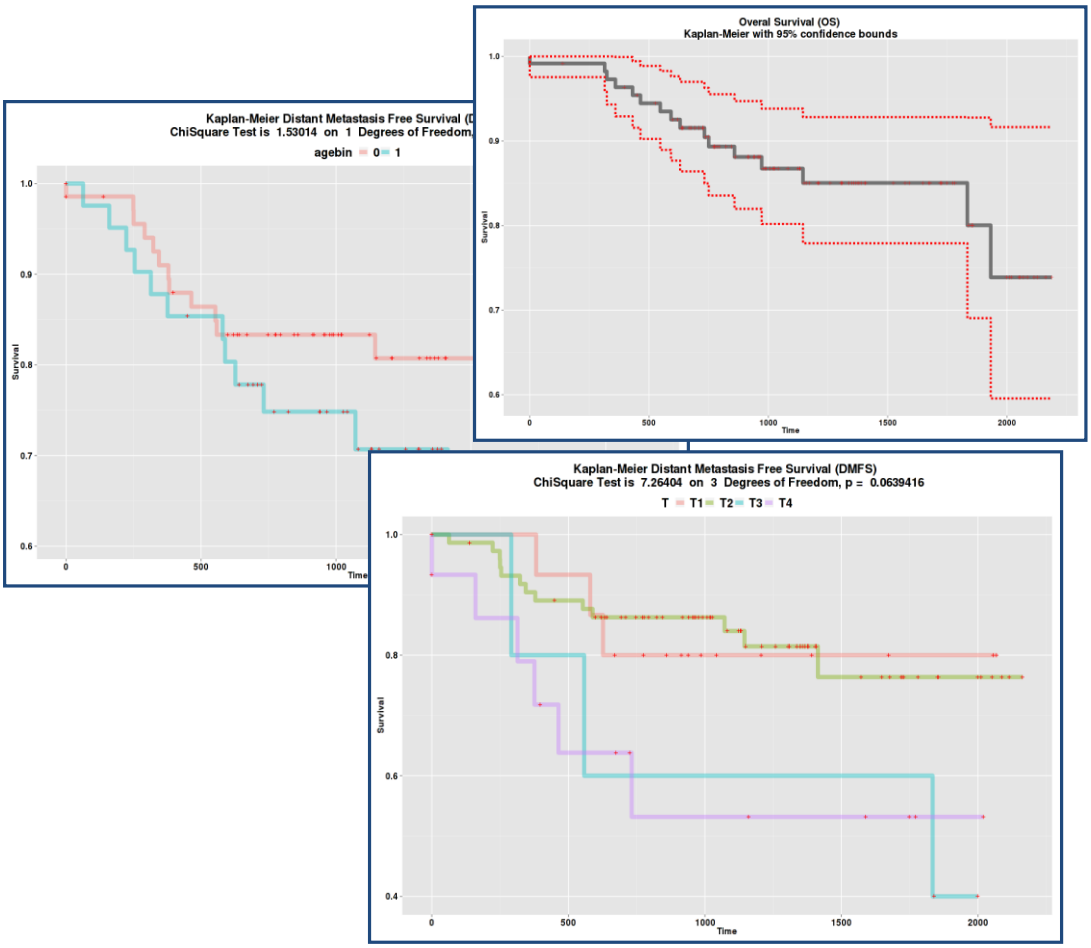

Figure 23 Survival Analysis applied in several cohorts

2.3.1.4 Performing quality control tests to the genomic data and identifying statistically significant genomic information through unsupervised learning

A complete gene expression analysis is implemented within the Analysis Framework, starting from the measured probe intensities and locations from a hybridized microarray that are stored in ".CEL" files to the identification of differentially expressed genes when clinical information such as the pathological complete response (pCR) is known. The overall analysis is divided into four main parts; read and normalize the signal intensities of the probes, the quality diagnostics, the expression measures, and the part which focuses on the assessment of expressed genes in discriminating clinical response. An overview of the genomic analysis framework is outlined in the following figure.

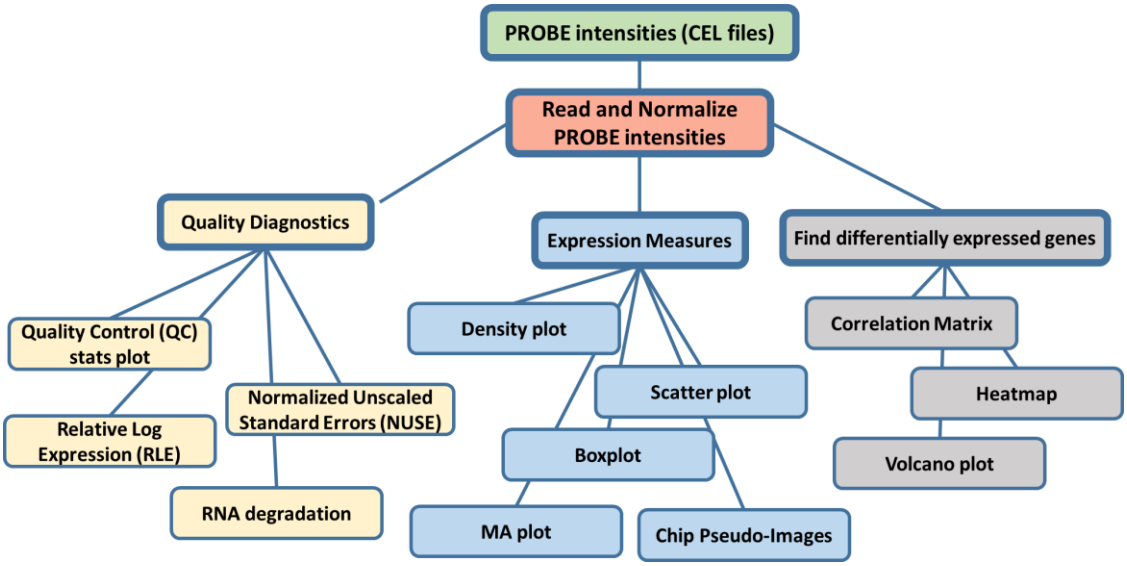

Figure 24 Overflow of the genomic analysis

At first, the raw information is stored in ".CEL" files and a number of pre-processing steps is required to retrieve it and produce gene expression estimates. These steps involving background correction, normalization, and summarization are combined into a single all-in-one pre-processing algorithm that takes raw probe intensities as input and produces gene expression estimates as output. Then, widely used quality assessment techniques such as the Relative Log Expression (RLE)[11], Normalized Unscaled Standard Errors (NUSE)[12], RNA degradation and Quality Control stats plot focus in whether a highly skewed genomic array will unduly influence initial normalization of the data and whether outlier arrays of gene expressions can be reliably identified (see Figure 25). Expression measures and their graphical output such as density plots, scatter plots, boxplots, MA plots and chip pseudo-images are then give a clear overview of the processed gene expression estimates.

[11] Bolstad BM, Collin F, Simpson KM, Irizarry RA, Speed TP. Experimental design and low-level analysis of microarray data. Int Rev Neurobiol. 2004;6:25–58
[12] McCall MN, Murakami PN, Lukk M, Huber W, Irizarry RA. Assessing affymetrix GeneChip microarray quality. BMC Bioinforma. 2011;6:137. doi: 10.1186/1471-2105-12-137

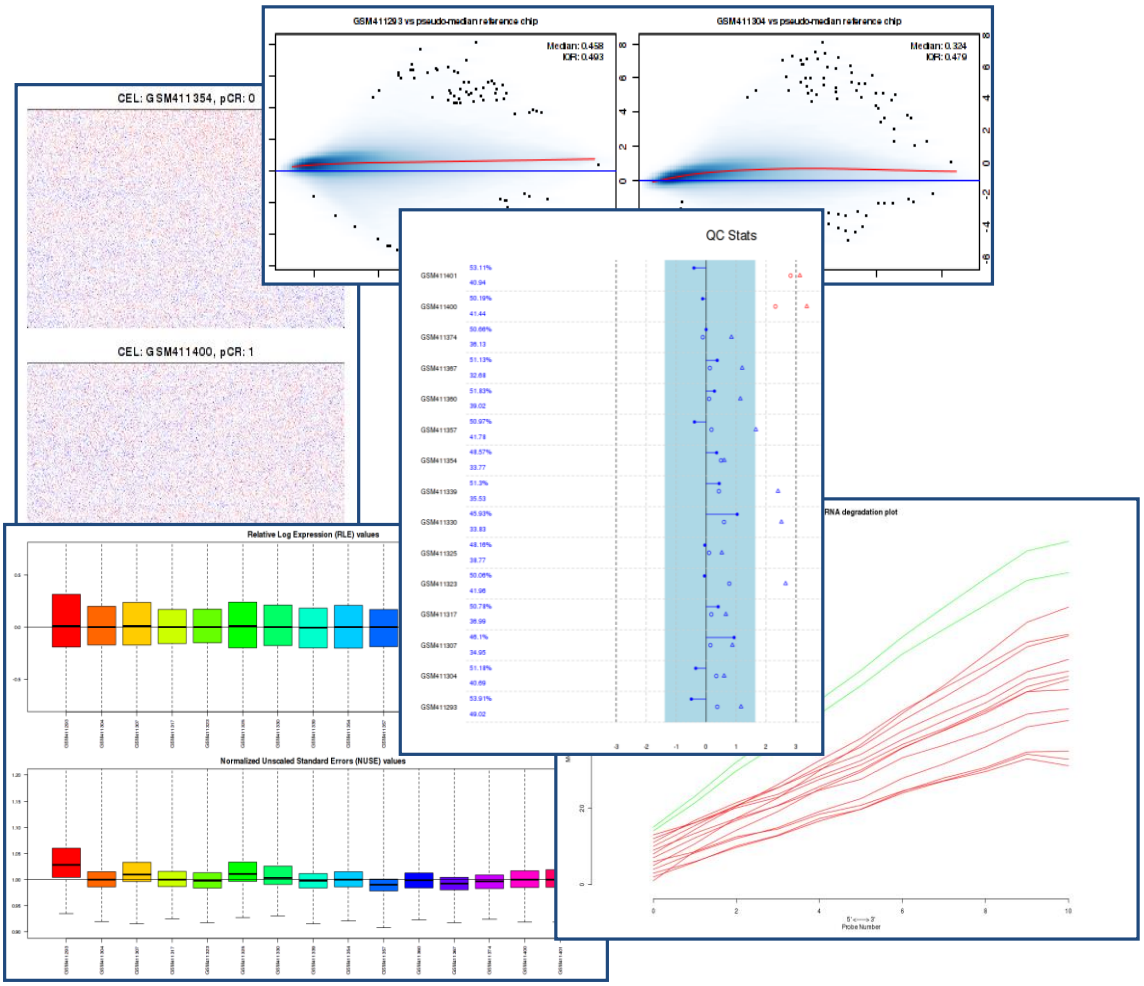

Figure 25 Quality diagnostics and visualization

A statistical analysis for differential expression of the genes is also implemented within the INTEGRATE Analysis Framework. This analysis involves a correlation matrix which is an all-by-all matrix of how well each sample's gene expression profile correlates with that of each other sample. A correlation coefficient is computed for each pair of arrays in the dataset and is presented qualitatively on a coloured matrix. The minimal value of this coefficient gives a good idea of the dataset homogeneity: low coefficients indicate important differences between array intensities.

A clustering heatmap is also produced where in the case of gene expression data, the colour assigned to a point in the heat map grid indicates the expression of a particular sample. The gene expression level is indicated by red for high expression and green for low expression. Coherent patterns of colour are generated by hierarchical clustering on both horizontal and vertical axes to bring like together with like. Cluster relationships are indicated by tree-like structures adjacent to the heat map, and the patches of colour may indicate functional relationships among genes and samples. At last, a type of scatter-plot named as volcano plot, is an effective and easy to interpret graph that summarizes statistical criteria like the t-test and the fold change in order to discount significant genes with misleadingly small variances and fold changes. The genes at the top of the graph are statistically the most significant and genes at the left

and right side of the graph have the largest fold-changes. Accordingly, the genes in the upper-right and in the upper-left corner are the most interesting genes as they show both a strong effect as well as high significance. An overview of this analysis is depicted in Figure 26.

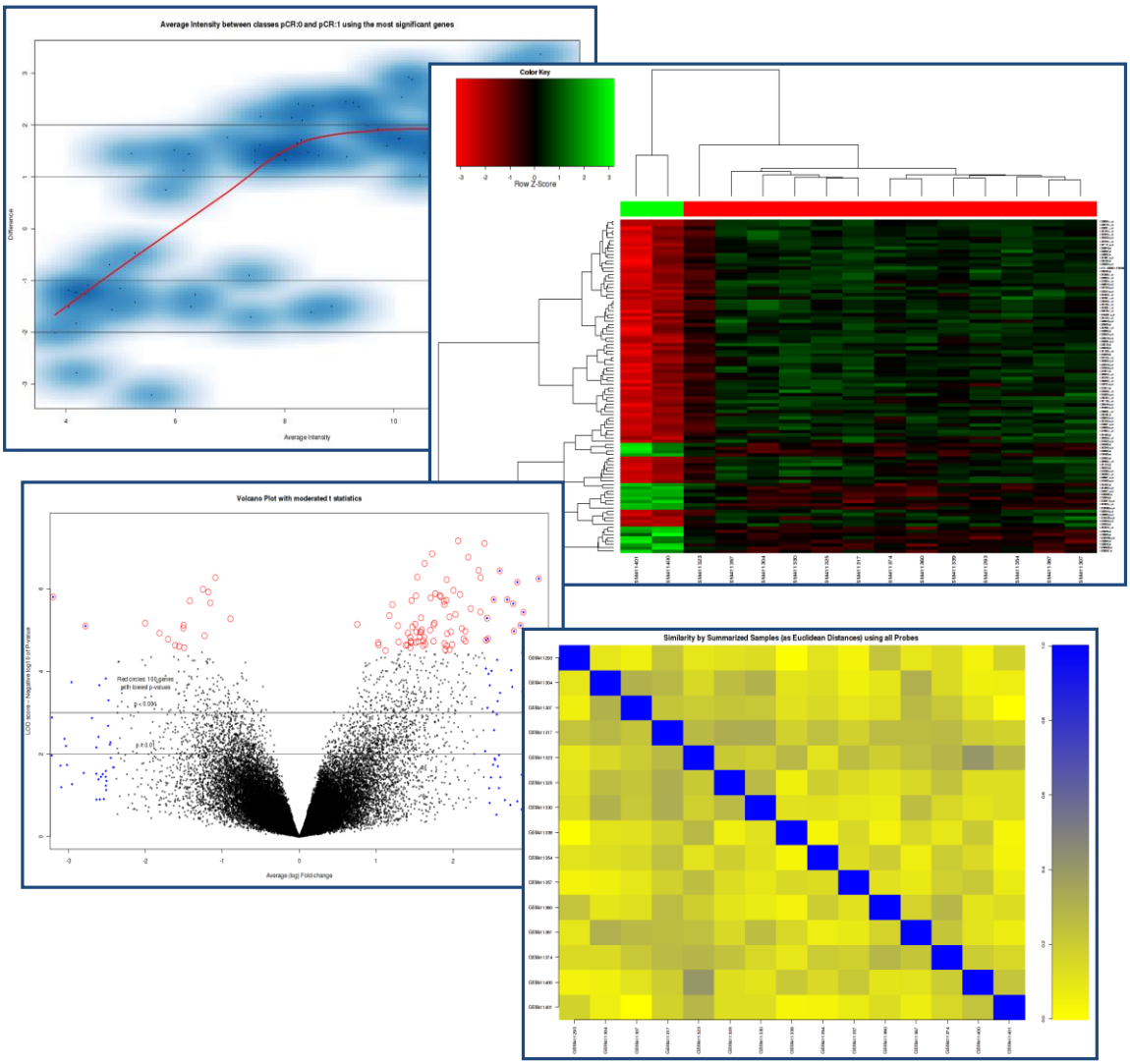

Figure 26 Highlighting any differential expression occurred to the genomic data

2.3.2 Using the Integrate Analysis Framework for Predictive Modelling

As mentioned in the previous topics, the overall INTEGRATE Analysis Framework is mainly divided into two broad categories; the statistical and the predictive analysis framework. The scenarios below highlight the need for prediction models that given a set of characteristics, predict in an accurate way the response to a drug X, the toxic effects of an investigational class of drugs, the response/resistance to a specific preoperative drug (i.e. epirubicin), and etc., using clinical characteristics such as the pathological complete response (pCR) of a patient. Biomedical data coming from different domains (e.g. microarray, clinical and proteomics) aim to provide enhanced information that leads robust operational performance (i.e. increased confidence, reduced ambiguity and improved classification) enabling evidence based management. Building a predictive model is not an easy task and a number of different techniques are incorporated including:

- feature selection methods for selecting a subset of relevant features from a large dataset that leads to better prediction than using the entire set
- pattern recognition methods for building the core functionality of the model
- data integration methods in case of using heterogeneous information to construct a model
- cross-validation methods for building unbiased models and assessing how the predictive results will generalize to an independent new data set
- statistical methods for evaluating the performance of the prediction model

Below, we present in a generic technical way the pipeline process followed in case a model is either constructed from homogeneous data or multi-modal data (heterogeneous integrated data).

2.3.2.1.1 Feature Selection Techniques

Feature selection (FS) techniques have become an apparent need in bioinformatics and in pattern recognition techniques. Specifically, the nature of microarray data poses a great challenge for computational techniques, because of their high dimensionality and their small sample sizes[13]. Therefore, combining predictive modelling and FS methods has become a necessity in many applications[14]. In both predictive models implemented within the INTEGRATE Analysis Framework, statistical feature selection methods, or alternatively filter-based FS methods, will first reduce the high dimensionality of the data before entering a model, and FS techniques embodied in the core functionality of the models with then assess the predictive power of the reduced dataset by assigning weight coefficients to each feature of the dataset.

[13] R. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics, 19(12):1484–1491, 2003

[14] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4):491–502, 2005

2.3.2.1.2 The Core Functionality of the Predictive Models

The predictive analysis framework of the INTEGRATE Analysis Framework relies on pattern recognition techniques and specifically on the Support Vector Machines[15] where mathematical equations build kernels from the data to be used as an implicit mapping of the input data into a high dimensional feature space (see Figure 27**Error! Reference source not found.**). This mapping, colloquially known as the "kernel trick" transforms observations with no obvious linear structure into observation easily separable by a linear classifier. This renders analysis of the data with a wide range of classical statistical and machine learning algorithms possible.

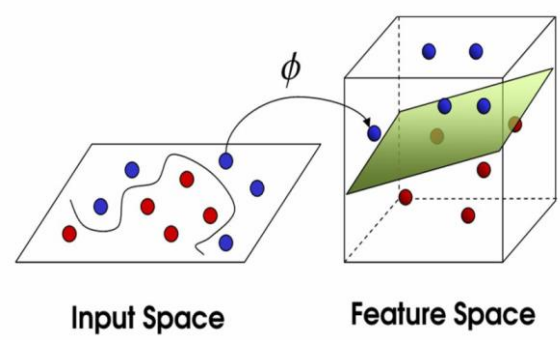

Figure 27 Principles of Kernel Methods depicting the "kernel trick"

2.3.2.1.3 Multi-modal Data Integration methods

In case of building a predictive model from homogeneous data, no data integration process is required and a single kernel is used. However, using a single kernel can be a limitation when the analysis requires integrated heterogeneous biomedical data from various data sources, since all features are merged into a unique kernel (i.e. features with their values ranging from 100 to 300 and features ranging from -1 to 1). To overcome this limitation, combining multiple kernels is necessary, like in the Multiple Kernel Learning (MKL) framework, pioneered by[16] to incorporate multiple kernels in predictive modelling.

The essence of MKL relies on the kernel representation while the heterogeneities of data sources are resolved by transforming each feature from the multi-modal dataset into kernel matrices. MKL involves first transforming each feature in a unique kernel framework ($K_{i,j}$), followed by weighted combination of the individual kernels ($d_{i,j}$) (see Figure 28). Following this methodology, we achieve building a predictive model using heterogeneous data and at the same time data integration is implemented where different data streams like clinical, microarray and multi-modal data in general are represented in a unified framework, overcoming differences in scale and dimensionality.

[15] Vapnik, V. The Nature of Statistical Learning Theory. Springer, N.Y. 1995
[16] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan. Learning the kernel matrix with semi-definite programming. Journal of Machine Learning Research, 5, 2004

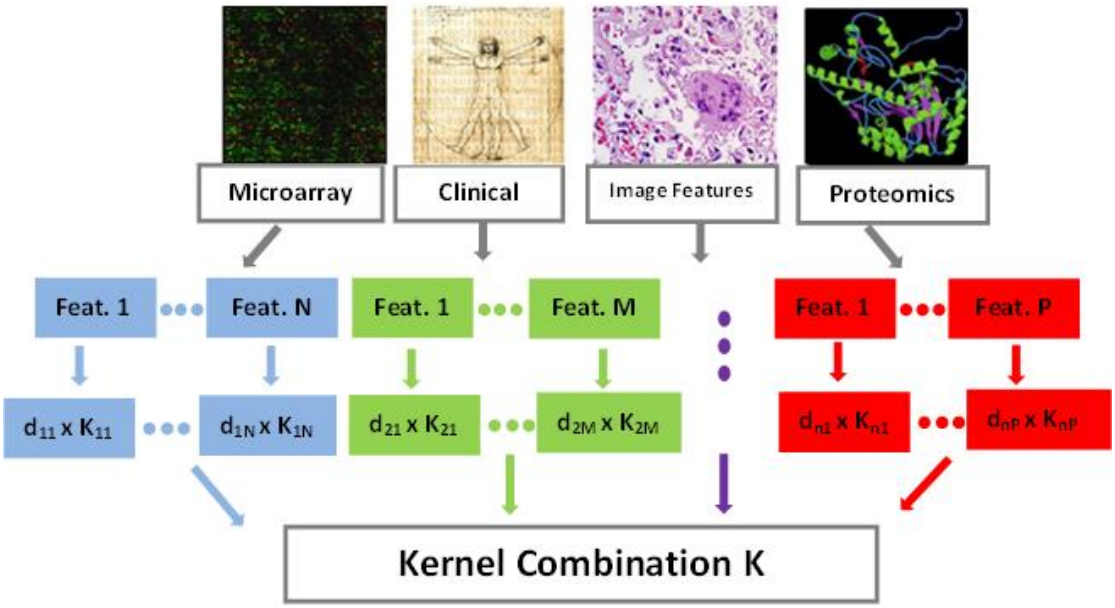

Figure 28 A schematic Representation of the Predictive Analysis Framework using Multi-Modal Data

2.3.2.1.4 Estimating the Generalization Error

In a predictive modeling framework, the goal is to build a model with good generalization. Such a model may demonstrate adequate prediction capability on the data used to build it (named also as training data) and on future unseen data (testing data). Cross validation is a procedure for estimating the generalization performance in this context in a way to protect the predictive model against over-fitting and it. In the INTEGRATE Analysis Framework, cross-validation will be run several times, increasing the number of estimates, where data is reshuffled before each run. Afterwards, several statistical measures for assessing the predictive accuracy of the model are computed and reported over the total number of the iterative procedure.

2.3.2.1.5 Evaluating the performance of the models

A crucial term for evaluation of classifiers is the classification error. However, in many applications distinctions among different types of errors turn out to be important. In order to distinguish among error types, several other statistical metrics like the Sensitivity, Specificity, Area under the Curve (AUROC), and several other display metrics (i.e. ROC and Precision-Recall curves) in corporation with the cross-validation techniques are used to evaluate the performance of the models.

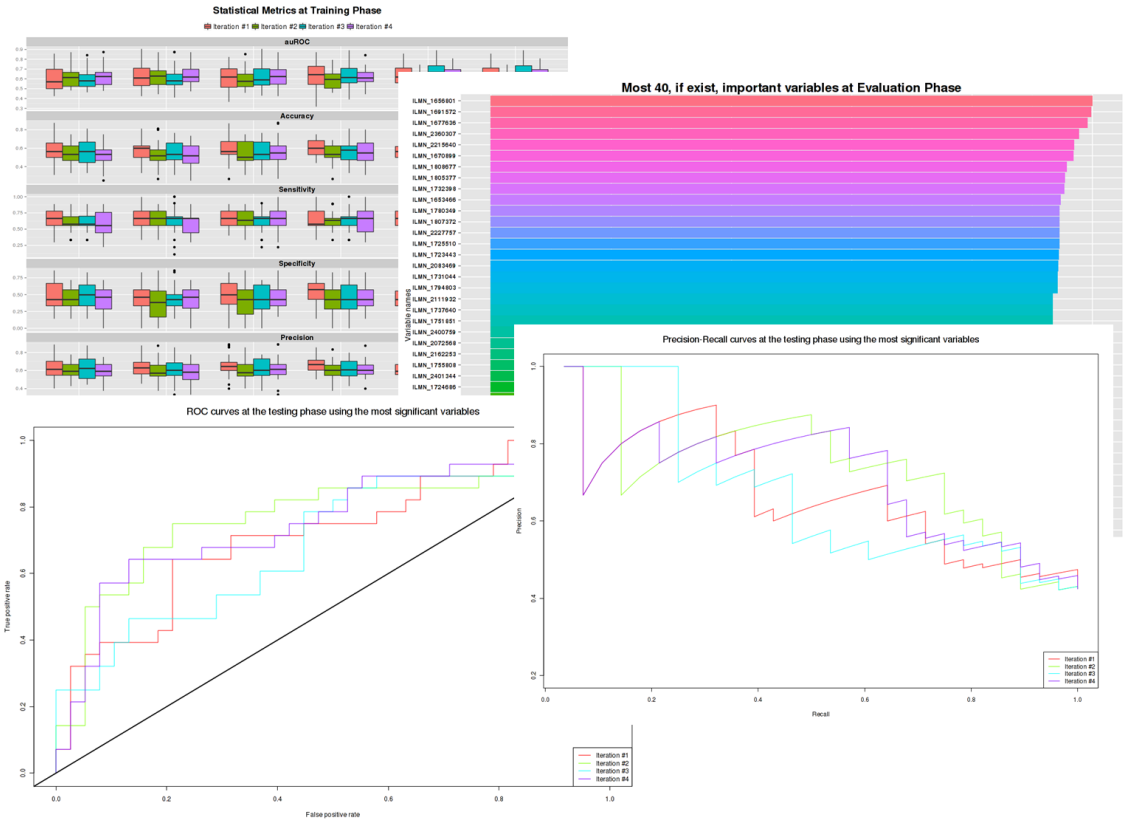

Figure 29 Indicative results of performance metrics for assessing the predictive accuracy of a model

3 SUMMARY

This deliverable presented the work described the implemented modelling framework and the predictive models for therapy response. The work addressed the main objectives WP5 regarding the approach and the methodology for developing multi scale predictive models in breast cancer and the construction of a corresponding framework for deriving such models based on the multi-level heterogeneous data provided by clinical trials in the neoadjuvant setting.

The deliverable focused on explaining the INTEGRATE Analysis Framework by first presenting the architecture and specifications and explaining the core functionality. Particular focus is given on presenting the portlets developed for the analytical tools and the predictive models as well as the integration with other tools within INTEGRATE. It is important to mention that the tools developed have been driven from Clinical Scenarios of the project (INTEGRATE VPH use in D.1.2) addressing statistical analysis and predictive modeling tasks.

The main goal of this work has been to empower scientists from diverse backgrounds to employ with ease (at the push of a button) sophisticated statistical analysis tools and to derive predictive models (again at the push of the button) from clinical trial data. To this end the framework has been designed in close collaboration with the Integrate consortium and in particular with the clinical site (BIG). It assists users in employing the statistical analysis tools implemented within the framework, addressing specific clinical questions and enables them to construct and validate their own predictive models within the context of BIG clinical trials. Once more, it is important to clarify that the clinical validation of any given model is out of the scope of this project and such validation will be organized at a later stage within the context of specialized trials within the participating clinical sites.

The main rationale for all this work has been the empowerment of users who wish to derive candidate biomarkers from large scale clinical trial data but are not able to employ sophisticated computational environments such as R and Matlab. Also, the Integration of this framework to the INTEGRATE environment means that all these tasks can be performed seamlessly and securely in a "one-stop shop" fashion.