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

Deliverable: D5.1 Report on the VPH use case study

Due date of deliverable: (9-30-2011)
Actual submission date: (10-7-2011)

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
PP	**Restricted to other programme participants (including the Commission Service**	
RE	**Restricted to a group specified by the consortium (including the Commission Services)**	
CO	**Confidential, only for members of the consortium (excluding the Commission Services)**	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
George Manikis	FORTH	
Kostas Marias	FORTH	
Manolis Tsiknakis	FORTH	

0.2 Documents history

Document version #	Date	Change
V0.1	30/6/2011	Starting version, template
V0.2		Definition of ToC
V0.3		First complete draft
V0.4	15/8/2011	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: George Manikis Partner: FORTH Address: N. Plastira 100, Vassilika Vouton Phone: +30 2810 391672 Fax: E-mail: gmanikis@ics.forth.gr
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer

Table of Contents

List of Figures

List of Tables

1 Definitions and Abbreviations

BIG	Breast International Group
pCR	Pathological Complete Response
VPH	Virtual Physiological Human
FDG	Fluorodeoxyglucose
PET	Positron Emission Tomography
GEO	Gene Expression Omnibus
ESR1	Estrogen Receptor 1
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
mAb	Monoclonal Antibodies
TKI	Tyrosine Kinase Inhibitors
HER	Human Epidermal Growth Factor Receptor
DFS	Disease Free Survival
OS	Overall Survival
CI	Confidence Interval
FISH	Fluorescent in situ Hybridization
OR	Odds Ratio
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
GEP	Gene Expression Profiling
SNP	Single Nucleotide Polymorphism
PCA	Principal Component Analysis
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
ROC	Receiver Operating Characteristic
AUC	Area under ROC curve
FS	Feature Selection
DEDS	Differential Expression via Distance Synthesis
SVM-RFE	Support Vector Machine-Recursive Feature Elimination
SVMs	Support Vector Machines
RBF	Radial Basis Function
ER	Estrogen Receptor

2 Introduction

Mathematical and computational modelling of cancer-related natural phenomena has been studied extensively over the last decades leading to a large number of either single scale or multi-scale models of cancer growth and/or response to therapy. The usual approach is the “bottom-up” approach i.e. starting from the molecular or cellular level and then trying to invoke higher levels. In addition to cellular proliferation and death which are at the core of most models, additional biological processes can be taken into consideration, including mutation and selection, angiogenesis [1] and invasion [2].

The Virtual Physiological Human (VPH) [3] is an initiative of the European Union that aims to support the development of integrative models of human physiology. Its central tenet is that fragmentation of research in physiology in different sub-disciplines is inefficient and ultimately does not allow building the realistic models that are needed in biomedicine. To be maximally useful, in silico physiological models have to be descriptive, integrative and predictive [4].

VPH-type models of human cancer can span several scales from the gene to the biological pathway, the cell, the tissue and finally the tumor in its environment. They take into account the three-dimensional organization of the tumor and its dynamics [5]. Building and validating integrative dynamical models of human cancer that encompass all the relevant biological processes is not yet feasible and only selected sub-systems are modeled. Moreover, it is difficult for technical and ethical reasons to obtain from human subjects the multi-scale repeated measurements that are needed, and parameters have been obtained mostly from model systems such as tissue culture, spheroids, or tumor xenographs.

Within INTEGRATE, we will initially focus on statistical models for cancer classification and for prediction of cancer prognosis and treatment response. These statistical models of cancer are very relevant in their own right from a clinical point of view. But they will also be useful for VPH-type modeling because they will provide clues about the identity of the relevant components and sub-systems. For example, the fact that a gene signature predictive of cancer prognosis incorporates an important immune component [6] suggests that a realistic physiological model of this type of cancer should incorporate this component.

Modeling at the molecular/genetic level aims to understand the cellular and genetic factors that play significant roles in oncogenesis and response to therapy (e.g. drugs). The research at this level takes into consideration key genes, cellular kinetics, pharmaco/ radiosensitivity dependence on the cell cycle phase etc. In this context, predicting therapy sensitivity from individual patient molecular profiles (e.g. microarrays) is a very challenging task [7]. At the tissue level the challenge is to simulate growth over time and response to various therapeutical regimes, aiming at the a priori definition of the optimal individual therapy for the patient [8-10]. The challenge in this field is the gradual coupling of models from various scales (related to the corresponding complex biological processes), which will lead to a better understanding of oncogenesis [11].

The main objectives of this work package (WP) are to propose an approach and a methodology and to build a framework enabling the development of multi-scale predictive models of response to therapy in breast cancer, making use of multi-level heterogeneous data provided by clinical trials in the neoadjuvant setting. The models developed in this work package (WP) will be based on realistic clinical research scenarios, in which have been developed based on the neoBIG research program, and on comprehensive data sets from rigorously conducted breast cancer clinical trials. The model-building tools may later be applied to other data sets, for example those resulting from prospective molecular screening, or from follow-on translational research studies using data and samples collected in the context of clinical trials. The models will also be used to validate the INTEGRATE approach and the appropriateness of the INTEGRATE infrastructure.

By proposing a methodology and building a framework for predictive models development within clinical trials we will support more efficient development and validation of such models and contribute to their faster adoption into clinical practice. We will make use of existing solutions, tools and standards whenever available and suitable for our scenarios. On the other hand, we will develop novel methods and computational approaches whenever existing methods evaluated as inadequate for the tasks at hand.

2.1 Breast cancer modelling and going beyond the state-of the art

The main modeling efforts related to breast cancer concern biostatistical models of risk of cancer, prognosis and relapse [12]. In the context of large scale clinical trials, prediction of outcome and individualization of therapeutic strategies are crucial when trying to improve prognosis and reducing patient suffering due to unnecessary treatment [13]. Therefore, a more realistic effort adopted within INTEGRATE is to exploit the unique opportunity of its NeoBIG empowered collaborative environment and combine multi-scale biomarkers (from genetic level to tissue level including imaging biomarkers) in order to define a methodology for improving the prognostic power of currently used practices for assessing neoadjuvant therapies. Figure 1 depicts the synergy between the BIG and NeoBIG research and Figure 2 shows the envisioned workflow of development and validation of predictive biomarkers in NeoBIG trials. This will eventually empower the clinician to predict/define early the responsiveness of the chosen chemotherapy regimens.

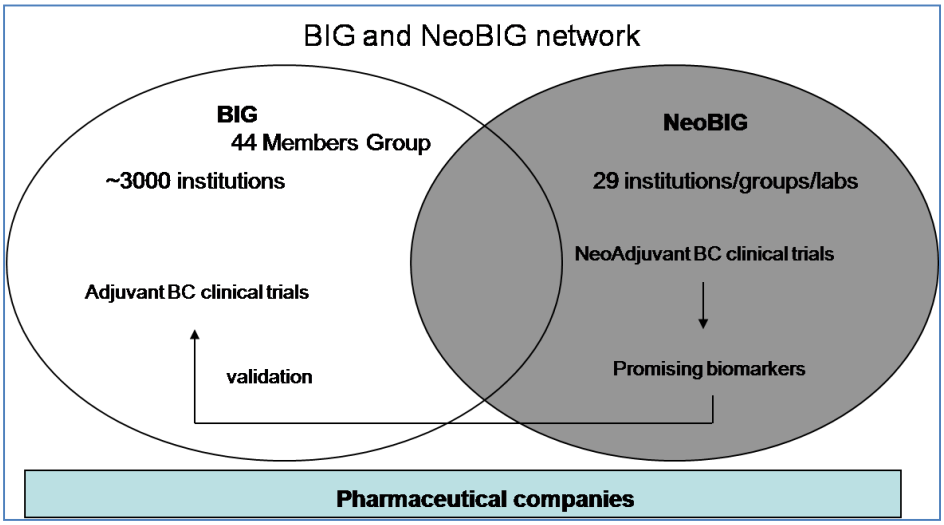

Figure 1 The synergy between BIG and NeoBIG

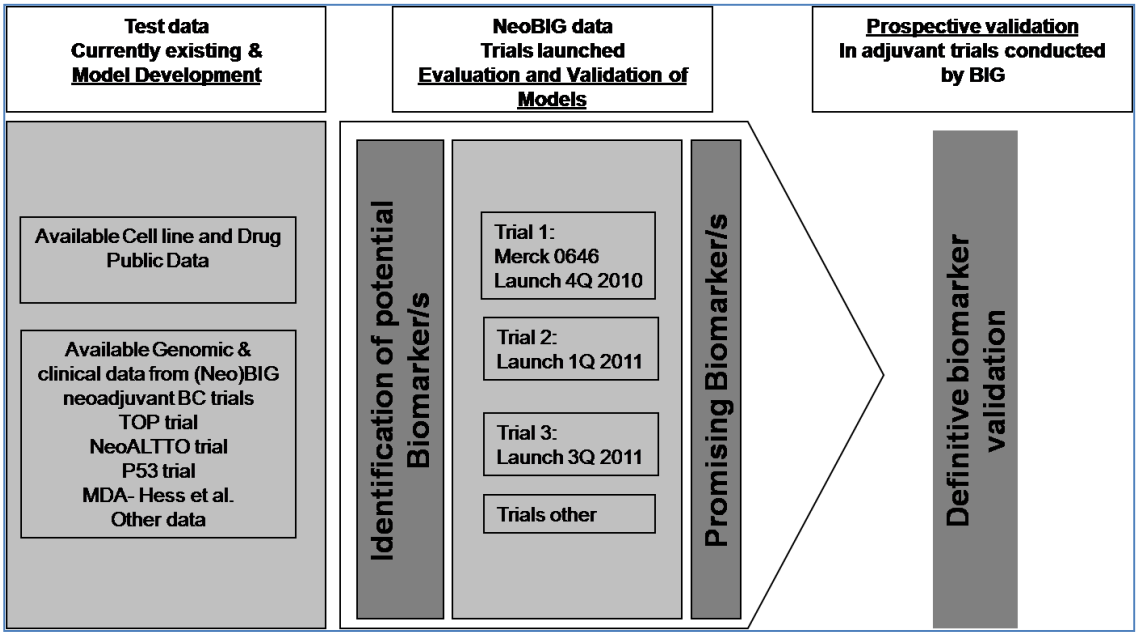

Figure 2 The Development and Validation of Predictive Biomarkers

The neoadjuvant setting, where therapy is administered prior to surgery, is a promising new arena for addressing many of the challenges in both clinical and translational research faced by clinicians today. There are a number of reasons and advantages for employing the neoadjuvant approach:

- Neoadjuvant systemic therapy produces outcomes equivalent to adjuvant systemic therapy, with an increased likelihood of breast conserving surgery and hence is a safe and viable option for breast cancer patients [14].
- Breast cancer is a common disease usually diagnosed in healthy women who do not have other co-morbidities that might preclude participation in clinical trials;
- The primary tumor is readily accessible for serial biopsies during treatment;

- Surrogate short-term endpoints such as pathological complete response rate (pCR) have been proven to be strongly predictive of long-term survival for treatment modalities such as chemotherapy and are rapidly available within a short time frame;

This allows for obtaining multiple serial biopsies and images, to characterize at biological multiple levels response to new agents. Furthermore, the existence of a surrogate clinical endpoint allows clinicians to rapidly evaluate if the new drug is more efficacious than the currently used standard of care ones.
This will take the form of a ‘use-case’ VPH scenario emanating from and being deployed within the INTEGRATE environment. The goal is to demonstrate that the predictive power of responsiveness can be enhanced by using multi-scale biomarker signatures.

3 SUMMARY

This report is based on some of the clinical scenarios elaborated so far in WP1, focusing on the VPH aspect of the project.

The report first summarises the multi-modal data that will be utilised in the context of developing predictive models. This is an ongoing effort for the project since it's crucial for developing predictive models. In this phase all data used will be retrospective data.

Then, clinically relevant questions are defined in the context of VPH predictive scenarios. The aim is to develop within the scenarios prediction models that given a set of characteristics will be able to predict in an accurate way the response to a drug and/or the response/resistance to a specific preoperative drug.

Last, the main techniques that will be exploited are reported in detail.

4 Data description

In this section, we describe the INTEGRATE data that will be used for cancer modelling. Data from the TOP clinical trial will be the first data to be shared on the INTEGRATE platform and use for modelling and thus we will start this section by describing them. After this, we will present other data types that are likely to be shared on the INTEGRATE platform and will be useful for cancer modelling.

4.1 Available data from TOP clinical trial

4.1.1 Clinical Data

These data are available for all patients from the TOP clinical trial. The clinical data, presented in Table 1, comprise information on tumour size, axillary lymph node status, tumor grade, biomarker expression status (estrogen receptor, progesterone receptor, HER2, TOP2A), and several clinical endpoints such as pathological complete response, distant metastasis-free survival and overall survival.

Variable	Supplementary Information
geo_accn	GEO accession numbers.
age.bin	$0: \leq 50$ years old, $1: > 50$ years old.
T	$T1: \leq 2cm$, $2cm < T2: \leq 5cm, T3: > 5cm$, $T4$:tumor of any size with direct extension to the chest wall or skin.
N	Axillary lymph node status: N0: no axillary lymph node metastasis, N1: metastasis in movable ipsilateral axillary lymph node(s), N2: metastasis in fixed ipsilateral axillary lymph node(s) or in clinically apparent ipsilateral internal mammary lymph node(s) in the absence of clinically evident axillary lymph node metastasis, N3=metastasis in ipsilateral infraclavicular lymph node(s) with or without axillary lymph node involvement; or in clinically apparent* ipsilateral internal mammary lymph node(s) in the presence of clinically evident axillary lymph node metastasis; or metastasis in ipsilateral supraclavicular lymph node(s) with or without axillary or internal mammary lymph node involvement.
Grade	Tumor grade (1, 2, 3)
HER2.bin	HER2 status by fluorescent in situ hybridization (FISH): 0: not amplified ($ratio < 2$), 1: amplified ($ratio \geq 2$).
TOP2A.tri	TOP2A status by FISH: -1: deleted ($ratio \leq 0.8$), 0: not amplified ($ratio < 2$), 1: amplified ($ratio \geq 2$).
topo.IHC	Topo by immunohistochemistry (%).
ESR1.bimod	ER status identified by bimodality of ESR1 gene expression.
ERBB2.bimod	HER2 status identified by bimodality of ERBB2 gene expression.
FINAL_ANALYSIS	Eligible patients included in the prediction analyses [15].
pCR	Pathological complete response. 0: no pCR, 1: pCR
DMFS_event	Distant metastasis free survival event.
DMFS_time	Distant metastasis free survival (days).
OS_event	Overall survival (event)

Table 1 Clinical TOP Trial dataset

4.1.2 Radiology Imaging Data

Mammography data (x-ray radiography of the breast) are available for a handful of patients from the TOP trial. The resolution of these images, stored in the DICOM format, is 70μm. They don't have associated annotations (e.g. tumour contours).

4.1.3 Genomic Data

4.1.3.1 Gene Expression Data

Affymetrix U133 plus 2.0 contains probes for more than 38,500 transcripts corresponding to well-characterized genes and Unigene genes, giving a full-genome view of gene expression. 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 often combined into a single all-in-one pre-processing algorithm that takes raw probe intensities as input and produces gene expression estimates as output.

4.1.3.2 Affymetrix SNP and CNV data

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation and represent over 80% of the genetic variation between individuals. SNPs are ideal candidates for research correlating phenotype and genotype. Since some SNPs predispose individuals to a certain disease or a trait or cause an altered reaction to a drug, they are proving to be highly useful in diagnostics and drug development. With more than 1.8 million genetic markers, Affymetrix' SNP 6.0 array provides high-performance, high-powered and low-cost genotyping. It is now available from Asuragen. In combination with Asuragen's service expertise you have the tools to carry out a whole-genome study and bring power to your research.

SNP array 6.0 contains probes for more than 906,600 single nucleotide polymorphisms (SNPs) and more than 946,000 probes for the detection of copy number variation (CNV). This corresponds to a median inter-marker distance in the genome of less than 700 nucleotides. Again, the analysis will start from the ".CEL" files, which allows maximum flexibility in the choice of the algorithms for CNV genotyping.

4.1.3.3 Illumina Methylation Data

This array allows interrogating the methylation status of 27,578 highly informative CpG sites located in the proximal promoters of 14,475 protein coding genes. This corresponds to an average of two interrogated CpGs per genes although a subset of more than 200 cancer-related genes has 3-20 interrogated CpGs. The Infinium assay uses a pair of probes for every CpG, with one probe measuring the level of the methylated CpG and the other probe measuring the level of the unmethylated CpG. The methylation of the CpG is then often expressed as a beta value, which is the ratio of the methylated signal on the sum of the methylated and unmethylated signal. Thus, beta values vary from 0.0 for a fully unmethylated CpG to 1.0 for a fully methylated CpG. These data are available for 34 patients from the TOP trial.

4.2 Expected data from other clinical trials

4.2.1 Radiology Imaging Data

For some of the trials, radiology images will be generated, in particular PET/CT images. PET-CT (Positron Emission Tomography – Computed Tomography) images are acquired in a device that combines detectors for the two modalities. The two images are then fused during co-registration. The FDG-PET part of the composite image allows the detection of anatomical regions with high metabolic activity, most prominently primary tumours and metastases, while the CT part of the composite image allows precise localisation of the anatomical structures and the tumours and metastases. PET/CT images are stored in the DICOM format. In some cases the contours of the primary tumour and other anatomical regions and landmarks of interest will have been delineated by a doctor (stored as a DICOM Structured Report).

4.2.2 Digital Pathology Images

Digital pathology images (scanned images of pathology microscope slides) will also be available on the platform and could be used for modelling. Many pathology slide scanners routinely used today have a magnification of 40X, although models with oil immersion of the objectives achieve a magnification of 100X.

Images with different techniques of tissue staining will be available. The most common stain in histology is the unspecific hematoxylin and eosin stain, which is suited to study the morphology of the cells and tissues. In addition to hematoxylin and eosin straining, immunohistochemistry will also be used. In this technique, antibodies binding to specific antigens in the tissue (e.g. a particular protein) are used to obtain a targeted colouring of the regions containing this antigen.

A large number of microscopy slide scanners exist, from different vendors, and the image file formats that they use are often proprietary. Many of these formats, however, are extensions of the TIFF image format with annotation metadata.

4.2.3 High-throughput Sequencing Data

A recent alternative to gene expression profiling with microarrays is *RNA-seq*, in which RNA is sequenced with one the new high-throughput sequencing (HTS) platforms. Typically, several hundred millions of short sequence reads are generated in such an experiment, which allows an unbiased estimate of the number of copies for each transcript. An advantage of RNA-seq compared to microarrays is that they can detect previously uncharacterized transcripts (small non-coding RNAs, microRNAs,…) because it doesn't rely on predefined sets of probes. Additionally, the sequence itself can be used to detect potentially oncogenic mutations or other functionally important sequence variants. Some complications with these data are their sheer volume and the relatively short length of the sequence that sometimes makes unambiguous mapping of their position in the genome impossible. Targeted sequencing is a related technique that uses selection of specific genomic regions or genes before sequencing, allowing focusing on these regions.

A representative platform for high-throughput sequencing is the Illumina HiSeq 2000 platform which can generate in about ten days up to 600GB of sequence data consisting of paired-end reads of 100bp.

5 Clinical Scenarios

The clinical scenarios that are going to be utilised in WP5 are given in the following chapters. In each chapter, the objectives, the steps required and the final results given by the examined scenario are briefly presented in a table format. A detailed presentation of the methodology required to achieve each scenario is also provided along with examples, template figures and tables.

5.1 Predictive Modelling Methodologies

The scenarios below highlight the need for a prediction model that given a set of characteristics, predicts in an accurate way the response to a drug X, the toxic effects of an investigational class of drugs and the response/resistance to a specific preoperative drug (i.e. epirubicin). 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 prediction model from different data sources is not an easy task. Its architecture is divided in several stages, including:

- Feature extraction from images.
- Feature selection methods for selecting a subset of relevant features.
- Data integration methods for constructing an informative meta-dataset.
- Building accurate classifiers for the prediction work.
- Pattern recognition methods for estimating the generalization error of the prediction model.
- Statistical methods for evaluating the performance of the prediction model.

The following chapters will guide the reader to a brief representation of the previously mentioned techniques before analysing our prediction models for the scenarios described below.

5.1.1 Feature Extraction from Images

The data generated by the omics and imaging technologies do not lend themselves to immediate incorporation in computational models of cancer but must be pre-processed or in some cases even extracted from the raw data.

Advances in image processing and computer vision nowadays allow the automated extraction of features from radiology and pathology images. While automated segmentation of radiology images cannot replace manual annotation by doctors, it can help them to delineate the three-dimensional shape of tumours efficiently. Similarly, automated algorithms are far from the reliability and expertise level of human pathologists, but they are already used to extract simple features from digital pathology images such as cell counts, biomarker quantification or basic morphological descriptors. The advantage of these automated software, is that, provided sufficient computational resources, they can process large areas the images. They are also unaffected by the biases linked to inter-observer variability.

5.1.2 Feature Selection

Feature selection (FS) techniques have become an apparent need in bioinformatics and specifically in pattern recognition techniques. Specifically, the nature of microarray and proteomic data poses a great challenge for computational techniques, because of their high dimensionality and their small sample sizes [16]. Many widely used methods were originally not designed to cope with large amount of irrelevant features. Therefore, combining pattern recognition techniques with FS methods has become a necessity in many applications [17]. In the current study, we focus on the supervised classification in which feature selection techniques can be organised into three categories; filter, wrapper and embedded techniques. An extensive overview of some of the most important feature selection techniques is given by [18].

Filter based techniques rely on information content of features. Different metrics from statistics like distance metric, information measure, correlation, etc. can be used to extract useful subsets from the entire dataset. In most cases a feature relevance score is calculated and low scoring features are removed. Advantages of filter techniques are that they easily scale to very high-dimensional data, they are computationally simple and fast, and they are independent from the classification procedure.

A novel technique for microarray feature selection called Differential Expression via Distance Synthesis (DEDS) will be adopted for the needs of our study [19]. This technique is based on the integration of different test statistics via a distance synthesis scheme because features highly ranked simultaneously by multiple measures are more likely to be differential expressed than features highly ranked by a single measure. The statistical tests combined are ordinary fold changes, ordinary t-statistics, SAM-statistics and moderated t-statistics. A recently published work that used DEDS technique can be found in [20], in which DEDS was applied in microarray data in order to reduce the high dimensionality of the dataset before contributing to the integrated meta-dataset for clinical decision support.

In general, classifiers cannot successfully handle high dimensional dataset generated from proteomics experiments. To overcome this problem, in case of proteomics, Wilcoxon rank sum test [21] as a feature selection scheme will be used to reduce the dimensionality of the proteomic dataset to a manageable number. Wilcoxon rank test is a nonparametric test which has no distribution assumption and when applied to the analysis of microarray data in [22], outperformed all other methods. All the data are ranked together based on their values. Then the ranks from one class are compared with those from the other class. A similar study is given by a biomedical data fusion framework in [20] that used this non-parametric rank test in proteomic data for extracting the most relevant proteins.

Therefore, a very first approach of feature selection will be implemented as a pre-processing step to reduce the high dimensionality of both microarray and proteomic data. DEDS and Wilcoxon rank test will be independent to the classification procedure, focusing exclusively to the reduce dimensionality, removing irrelevant and redundant data and improve discrimination between the examined classes. The idea behind applying filtering techniques is that we want to avoid time consuming feature selection techniques keeping at the same time unbiased the classification approach that will be implemented at the next step. The next step, described by the following chapter, is the integration of the different data sources into a unique meta-dataset.

5.1.3 Integrating Heterogeneous Data

5.1.3.1 Integration of Genomic Data

Integration of multiple types of genomic data can produce high-quality predictive models and shed new light on the molecular mechanisms at play (Cancer Genome Atlas Research Network, 2011). This cannot be achieved by simply piling up the data but data needs to be integrated. Multiple mechanisms for multi-level genomic data integration are possible.

The first level of integration of genomic data is identifier mapping. For example, the oligonucleotide probe set detecting a particular transcript on a gene expression microarray must be linked to the name of the corresponding gene. Similarly, identifiers for CpG sites on a DNA methylation microarray or for probes on a CNV microarray must be linked to the names of the corresponding genes. Although tools and databases exist for this purpose, it is not trivial as there is rarely a one-to-one unambiguous mapping between different molecular entities and the corresponding identifiers.

At a higher level, molecular pathways provide a powerful unifying framework for genomic data integration. Disturbances over sets of genes that do not make sense when they are considered individually become meaningful when these genes are mapped to biological pathways.

Finally, integration at the level of the biological functions themselves can bring insight and clarity, for example through the use of ontologies such as GeneOntology.

5.1.3.2 Machine Learning Methods for Integration

In addition to integration methods specific of genomic data, generic methods for the integration of high-dimensional multi-level data sets have been developed in recent years, especially with the machine learning community. We present some of these methods here.

With a wide array of multi-modal and multi-scale biomedical data available for disease characterization, the integration of heterogeneous biomedical data in order to construct accurate models for predicting diagnosis, prognosis or therapy response seems to be one of the major challenges for the data analysis. Different data streams like clinical information, microarray and proteomic data will be represented in a unified framework, overcoming differences in scale and dimensionality. Data integration, or alternatively data fusion, is a challenging task and approaches for data integration like bagging, boosting [23] and Bayesian networks [24] allow different strategies to integrate heterogeneous data. These methods either use direct or indirect ways (i.e. at the decision level) of combing heterogeneous data. In this work, we formulate the data integration task in machine learning terms, and we rely on kernel-based methods to construct integrated meta-datasets for prediction analysis. During the last decade, kernels have been developed significantly because of their ability to deal with a large variety of data, for example Support Vector Machines (SVMs) [25], Kernel-PCA [26] or Kernel Fisher Discriminant [27]. Kernels [28] use an implicit mapping of the input data into a high dimensional feature space defined by a kernel function; a function returning the inner product $\langle \Phi(x), \Phi(x') \rangle$ between two data points x, x' in the feature space (see

Figure 3). More precisely, the dot product $\langle \Phi(x), \Phi(x') \rangle$ can be represented by a kernel function K as:

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

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. Any symmetric, positive semi-definite function is a valid kernel function, resulting in many possible kernels, e.g. linear kernel, Gaussian radial basis function (RBF) and polynomial (see following equations). Parameter σ stands for the tuning parameter of the RBF kernel, the scaling parameter, $scale$, of the polynomial kernel is a convenient way of normalizing patterns without the need to modify the data itself, and $degree$ is the degree of the polynomial. They all correspond to a different transformation of the data, meaning that they extract a specific type of information from the dataset.

$$K(x, x') = \langle x, x' \rangle, for\ Linear\ Kernel$$
$$K(x, x') = exp(-\sigma \cdot \|x - x'\|^2), for\ RBF\ Kernel$$
$$K(x, x') = (scale\langle x, x' \rangle + offset)^{degree}, for\ Polynomial\ Kernel$$

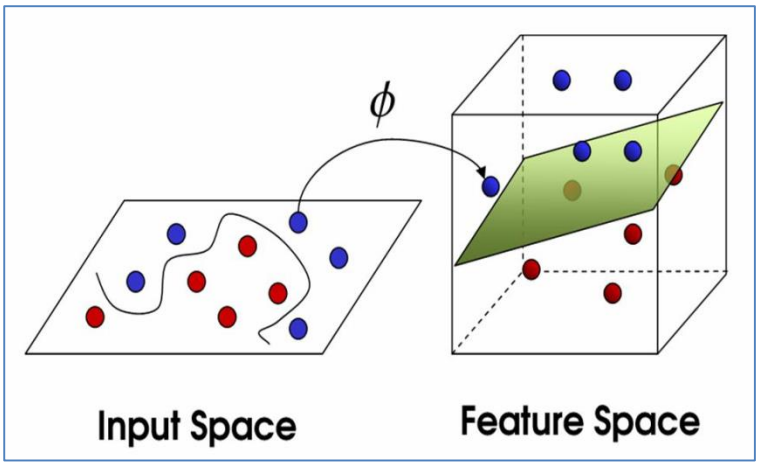

Figure 3 Principles of Kernel Methods

However, using a single kernel can be a limitation for some tasks (e.g. integrating heterogeneous biomedical data form various data sources), since all features are merged into a unique kernel. To overcome this limitation, combining multiple kernels is necessary, like in the Multiple Kernel Learning (MKL) framework, pioneered by [29] to incorporate multiple kernels in classification. The essence of MKL relies on the kernel representation while the heterogeneities of data sources are resolved by transforming the different data sources into kernel matrices. MKL involves first transforming each data source (e.g. clinical, microarray and proteomic data) in a common kernel framework, followed by weighted combination of the individual kernels as given by the following equation. M is the total number of kernels, each basis kernel K_m (i.e. linear, RBF or polynomial) may either use the full set of each data source or each feature from all datasets individually and the sum of the weighting coefficients d_m equals to

one. This approach has been proposed to tackle the descriptor fusion problem, by merging in a single kernel a set of kernels coming from different sources.

$$K(x, x') = \sum_{m=1}^{M} d_m K_m(x, x'), with\ d_m \geq 0, \sum_{m=1}^{M} d_m = 1$$

A graphical representation of the MKL approach is depicted in the following figure. The top schema (a) presents the MKL in which each basis kernel has been computed from the entire data source, respectively. Then, using the MKL methodology a combination of base kernels is computed. A slightly different approach is given in b) where a basis kernel is computed for each feature followed by a weight coefficient. A more detailed representation of the MKL methodology will be given in the following chapter in which multiple kernel learning is embodied into the classification task. It is highly important to mention here that when dealing with several data types of a specific group of data (i.e. two different microarray analysis datasets), a basis kernel is computed for each data type. For instance, in case we have gene expression (GE), single nucleotide polymorphism (SNP) and methylation data, for an analysis as in Figure 4 a), a basis kernel is computed for GE, SNP and methylation data respectively.

Given an introduction of both single and multiple kernel methodologies, we can now represent the main categories for data integration using kernels. Three ways exist to learn simultaneously from multiple data sources with kernel methods: early, intermediate and late integration [30]. In early integration, heterogeneous data are considered as one big dataset. A single kernel maps the dataset into the feature space and a classifier (e.g. Support Vector Machine) is trained directly on the single kernel. In intermediate integration, a kernel is computed separately for each homogeneous dataset, or for each feature of the datasets. Each of the kernels is given a specific weight, a linear combination of the multiple kernels is performed and a classifier is trained on the explicitly heterogeneous kernel function. At late integration, for each dataset (e.g. clinical, microarray and proteomics) a kernel is computed and a classifier is trained. The multiple outcomes of all the classifiers are combined with a decision function to become a single outcome.

In this work, we will perform intermediate integration because this type of data integration seemed to perform better on some genomic data sets [30]. Intermediate integration has the advantage that the nature of the data is taken into account when compared to early integration. On the other hand, when compared to late integration, intermediate has the advantage that a model is trained by weighting both datasets simultaneously through the use of kernels, leading to one decision result.

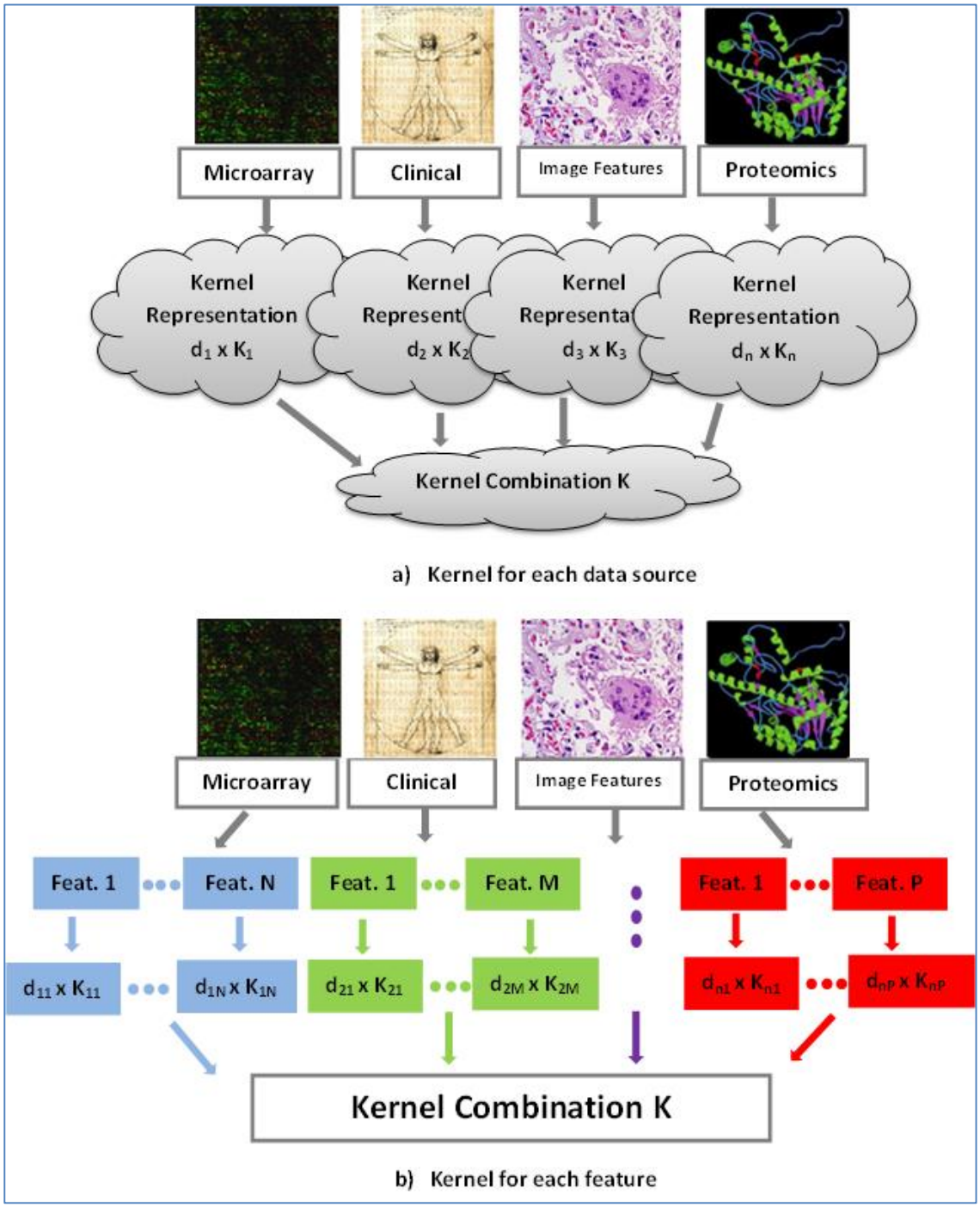

Figure 4 Multiple Kernel Learning

5.1.4 Kernel-Based Classification and MKL

The notion of Multiple Kernel Learning is originally proposed in a binary Support Vector Machine classification [25]. The SVM forms a linear discriminant boundary in kernel space with maximum distance between samples of the two considered classes. Among all linear discriminant boundaries separating the data, also named as hyper-planes, a unique one exists yielding the maximum margin of separation between the classes [31], as depicted in Figure 5.

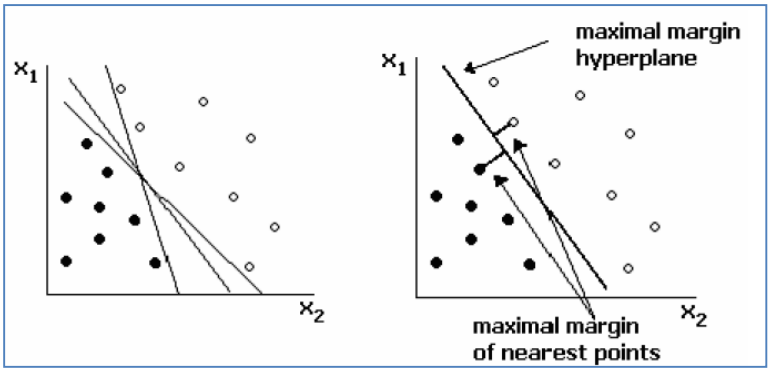

Figure 5 Linear Classification example [31]

Since SVMs are large margin classifiers, they have the potential to handle large feature spaces and prevent over-fitting [32]. Therefore, this methodology will be adopted in our study to handle the high dimensionality of the genomic data and perform the classification analysis. By replacing the single kernel with a combination of base kernels, the methodology is switched from the single kernel-based classification to the multiple kernel learning.

5.1.5 Decision Trees and Ensembles of Trees

Besides the kernel-based classification approaches, a second option for building our prediction models is given by the ensemble classifiers that consist of Decision Trees. In recent years, the ensemble classifier techniques are rapidly growing and enjoying a lot of attention from pattern recognition and machine learning communities due to their potential to greatly increase prediction accuracy of a learning system. These techniques generally work by means of firstly generating an ensemble of base classifiers via applying a given base learning algorithm to different permutated training sets, and then the outputs from each ensemble member are combined in a suitable way to create the prediction of the ensemble classifier. The combination is often performed by voting for the most popular class. Examples of these techniques include Bagging [33], AdaBoost [34], Random Forest [35] and Rotation Forest [36]. Among these methods, AdaBoost has become a very popular one for its simplicity and adaptability [37, 38].

AdaBoost constructs an ensemble of subsidiary classifiers by applying a given base learning algorithm to successive derived training sets that are formed by either resampling from the original training set or reweighting the original training set according to a set of weights maintained over the training set. Initially, the weights assigned to each training instance are set to be equal and in subsequent iterations, these weights are adjusted so that the weight of the instances misclassified by the previously trained classifiers is increased whereas that of the correctly classified ones is decreased. Thus, AdaBoost attempts to produce new classifiers that are able to better predict the ‘‘hard” instances for the previous ensemble members.

Based on Principal Component Analysis (PCA), a new ensemble classifier technique named Rotation Forest was recently proposed and demonstrated that it performs much better than several other ensemble methods on some benchmark classification

data sets [36]. Its main idea is to simultaneously encourage diversity and individual accuracy within an ensemble classifier. Specifically, diversity is promoted by using PCA to do feature extraction for each base classifier and accuracy is sought by keeping all principal components and also using the whole data set to train each base classifier. A possible decision tree construction for every ensemble classifier will be C4.5 Decision Tree [39].

5.1.6 Evaluating the performance of the classifier

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, a confusion matrix (see Table 2) can be used to lay out the different errors. In case of a binary classification problem, a classifier predicts the occurrence (Class Positive) or non-occurrence (Class Negative) of a single event or hypothesis.

	True Class	
Predicted Class	Class Positive	Class Negative
Prediction Positive	True Positives (TP)	False Positives (FP)
Prediction Negative	False Negatives (FN)	True Negatives (TN)

Table 2 Confusion matrix for classification

Common metrics for evaluation of the classification performance, calculated from the confusion matrix, are the sensitivity, specificity and accuracy. Using the notation in Table 2, these metrics can be expressed as:

$$Sensitivity = \frac{TP}{TP + FN} = True\ Positive\ Rate$$

$$Specificity = \frac{TN}{TN + FP} = True\ Negative\ Rate$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

In case where the number of True Positives is small when compared with True Negatives, precision can be also calculated.

$$Precision = \frac{TP}{TP + FP}$$

Kappa error or Cohen's Kappa Statistics [40] value will be used to compare the performance of the classifiers as well. Kappa error is a good measure to inspect classifications that may be due to chance. In [41] an attempt was made to indicate the degree of agreement that exists when the Cohen's kappa is found to be in various ranges; ≤ 0 (poor); $0 - 0.2$ (slight); $0.2 - 0.4$ (fair); $0.4 - 0.6$ (moderate); $0.6 - 0.8$ (substantial); $0.8 - 1$ (almost perfect). As the Kappa value calculated for classifiers approaches to 1, then the performance of the classifier is assumed to be more realistic rather than by chance. Therefore, in the performance analysis of classifiers, Kappa error is a recommended metric to consider for evaluation purposes [42] and it is calculated with the equation below.

$$Cohen's\ Kappa = \frac{\left[(TP+TN) - \left(\frac{\left((TP+FN)(TP+FP)+(FP+TN)(FN+TN)\right)}{n}\right)\right]}{\left[n - \left(\frac{\left((TP+FN)(TP+FP)+(FP+TN)(FN+TN)\right)}{n}\right)\right]}$$

Sensitivity, specificity and accuracy describe the true performance with clarity, but failed to provide a compound measure for the classification performance. This measure is given through Receiving Operating Characteristic (ROC) analysis. For a two-class classification problem ROC curve is a graphical plot of the sensitivity vs. 1-specificity as the discrimination threshold of the classifier is varied (see Figure 6).

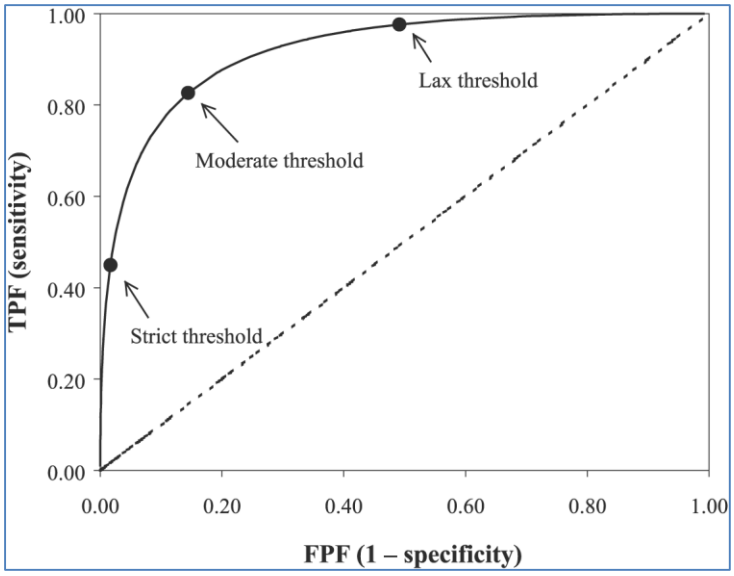

Figure 6 A typical ROC curve, showing three possible operating thresholds

While the ROC curve contains most of the information about the accuracy of a classifier through several values of thresholds, it is sometimes desirable to produce quantitative summary measures of the ROC curve. The most commonly used quantitative measure is the area under the ROC curve (AUC). AUC is a portion of the area of the unit square, ranging between 0 and 1, and is equivalent to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

Another useful plot diagnostic of model performance related to the ROC curve is the precision-recall curve [43], where recall is given by:

$$Recall = \frac{TP}{TP+FN}$$

5.1.7 Estimating the generalization error

In pattern recognition, a typical task is to learn a model for the available data. In a general classification problem, the goal is to learn a classifier with good generalization.

Such a model may demonstrate adequate prediction capability on the training data and on future unseen data. Cross validation is a procedure for estimating the generalization performance in this context in a way to protect the classification model against over-fitting. No matter how sophisticated and powerful algorithms for classification are developed, if no reliable performance estimates are obtained, no reliable decisions can be made based on classification results. Basic forms in cross-validation are the k-fold and the leave-one-out cross-validation.

In k-fold cross-validation the data is first partitioned into k equally (or nearly equally) sized folds. Subsequently k iterations of training and validation are performed such that, within iterations, a different fold of the data is held-out for validation while the remaining k-1 folds are used for learning. If k equals the sample size, this is called the leave-one-out. In this study, k-fold cross-validation (with k=5 or k=10) or leave-one-out, in case of few samples, will estimate the performance of our model. In case of k-fold cross-validation, the data will be stratified prior to being split into k folds in order to ensure that each fold is a good representative of the whole. Finally, stratified k-fold cross-validation will be run several times, increasing the number of estimates, where data is reshuffled and re-stratified before each run.

Conclusively, the generalization error will be estimated by applying extensive iterative internal validation using cross-validation techniques. K-fold and leave-one-out cross-validation allow each subset/sample to serve once as a test set, producing different measurements. Therefore, the means and standard deviation of the sensitivity, specificity, accuracy, precision and AUC will be computed and reported over the total number of the iterative procedure.

5.1.8 Feature Selection in Kernel Space

The MKL approach can be also extended in feature selection techniques applied to kernel space, where features that contribute to the highest discrimination between the classes are chosen as the most significant for classification [44-46]. Existing methods typically approach this type of problem as solving a task of learning the optimal weights for each feature representation. More specifically, for feature selection in a multi-dimensional space, MKL uses each feature to generate its corresponding kernel and aims to select the relevant features of the corresponding base kernels according to their relevance to the task of classification. In this way, the feature weights and the classification boundary are trained simultaneously and the most relevant features (features with the highest weighted value) that leading to the best classification performance are selected.

An alternative way of selecting the most relevant features in the kernel space is given in [47]. The heterogeneous data sources are integrated into a unique kernel framework, and the combined kernel matrix extracts the data in the form of pairwise similarities (or distances) which can be used as the input for a generic feature selection algorithm. Generally speaking, the features in the kernel space are not assumed to be independent. Therefore, feature selection methods that consider each feature individually are unlikely to work well in a kernel space. However, a margin-based feature selection method can handle the feature-dependency problem successfully, as explored in [48]. For that reason, methods like Relief [49] and Simba [48] can be adopted as a margin-based feature selection method. Simba is a recently proposed margin based feature selection approach, which uses the so-called large

margin principle [25] as its theoretical foundation to guarantee good performance for any feature selection scheme which selects small set of feature while keeping the margin large. Roughly speaking, the main idea of Simba is to obtain an effective subset of features such that the relatively significant features have relatively large weights by using hypothesis-margin criterion.

5.2 Scenario A-Retrospective use of data

Scenario A	
Objective	An academic researcher wants to define if the response to a specific drug X used across multiple breast cancer neo-adjuvant trials can be predicted by a gene expression signature.
Steps	• The researcher logs into the system. • The researcher filters by type of cancer (i.e. breast), the treatment setting (i.e. neoadjuvant) and the selected drug (i.e. drug X). • The academic researcher selects for the following outputs; gene expression data, pathologic response, trial name and additional characteristics. • The researcher either downloads the results on his computer (i.e. an excel file in csv format) and the gene expression data in the relevant format or works directly on the INTEGRATE platform using the provided tools.
Results	The researcher tries now to validate the predictive role of the gene signature using publicly available gene expression data generated from trial using the same drug X.

Table 3 Scenario A-Retrospective use of data

The objective of this study is to build a prediction model that given a set of characteristics, predicts in an accurate way the response to a drug X. Summarizing all the previously analysed techniques in chapter 5.1, we can now proceed to the presentation of our methodology for identifying the most relevant biomedical data that characterize in an accurate way the response of a drug X. Initially, the researcher logs into the INTEGRATE platform and exports the examined dataset which consists of patients with breast cancer, treated by any possible type of regimen under neoadjuvant therapy. All available patients are dichotomised into two classes based on their pathological response (pCR) to drug X. The multisource dataset might be include clinical, microarray and/or proteomic data. The available data enters our prediction modelling system and a pipelining approach, as presented in Figure 7, is executed.

Due to the very high dimensionality of the microarray and proteomic data, the first step is to perform a filter-based feature selection technique using DEDS and Wilcoxon rank test to microarray and proteomic data, respectively. In a problem with over 1000 features, filtering methods like Wilcoxon and DEDS have the key advantage of significantly small computational complexity contributing to a flexible dataset of about 100-200 features that enter the prediction model for further analysis. Our main aim in this step is to provide datasets with a manageable number of features for further analysis and not to thoroughly search for the best subset of features that lead to the best prediction accuracy.

The datasets with reduced dimensionality are next entering the data integration model. In multiple kernel learning, a basis kernel multiplied by a constant weighted value is

assigned to each feature from each dataset and a convex combination of the basis kernels is constructed. The patterns of the data in the form of pairwise similarities are extracted by the combined kernel matrix, which can then be used as the input for a generic kernel-based feature selection algorithm. The learning process is therefore executed into a kernel-based classification approach using SVMs for estimating the weights d for the base kernels and the parameters of the classifier. Alternatively, ensemble classifiers using Decision Trees in which a feature selection technique is embodied into the overall method (5.1.5) will be used as well.

The optimization problem follows a recursive procedure either by iterated stratified k-fold cross validation or leave-one-out (see chapter 5.1.7 for further details) split the overall dataset into training, validation and testing set. Following the kernel-based method, through the iterative procedure the weight coefficients of the basis kernels and the margin-based methods described in 5.1.8 define a small set of relevant features that contribute to the highest performance of the classifier. On the other hand, ensemble trees define their own relevant subset of features.

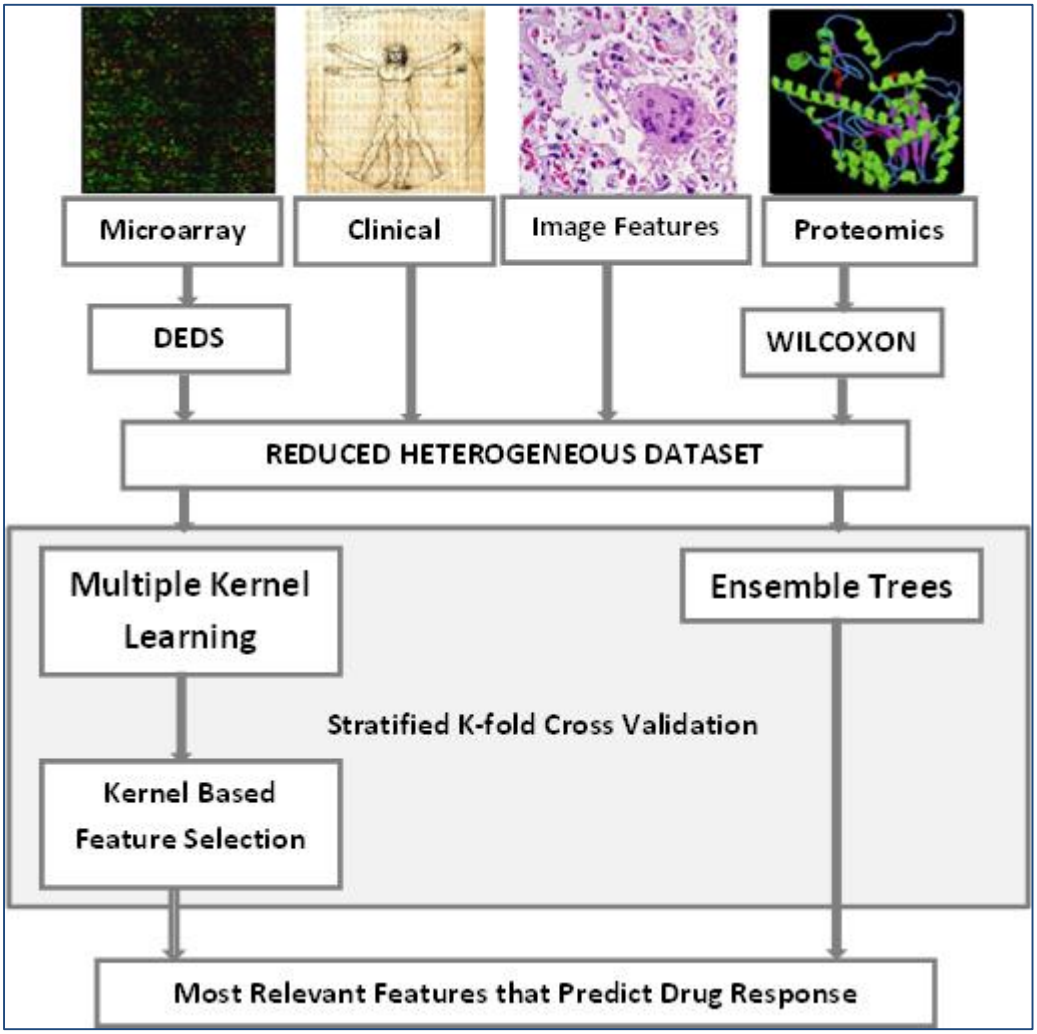

Figure 7 Overall framework for Scenario A

As we applied an iterative procedure for identifying the most relevant subset of features we actually obtained different selected subsets. By computing the frequency

of each feature appearing in all the subsets, we can identify and rank the important features which have been most frequently selected for different re-sampling sets. This is important because the most important features from a statistical point of view are also likely to be the most important from a biological point of view. Concerning the consistency of the classifier in terms of selected features as the most relevant over the iterative procedure, the consistency (or features overlap) index is tabulated in Table 4.

The model returns a matrix with the most relevant features along with their frequency of appearance through the iterative classification procedure, their ranking position in each iteration based on the kernel-based feature selection, a short statistical analysis, and a p-value. For instance, in case of a 10-fold cross-validation where 10 folds contribute to the estimation of the generalization error of the classifier the matrix is given as in Table 4. Rank ordered by t-statistics among the features of the integrated dataset will be provided as well. All p-values will be two-sided with statistical significance evaluated at the 0.05 alpha level.

Feature	Frequency of Appearance	Ranking Position	T-test	P-value
Feature 1	8 / 10	1-10-1-1-1-3-2-6		
Feature 2	6 / 10	4-2-8-10-6-1		
.	.	.		
.	.	.		
.	.	.		
.	.	.		
.	.	.		
.	.	.		
Feature M	2 / 10	.		

Table 4 T-statistics, ROC analysis, ranking of the selected features

In order to evaluate the potential of the classifier to discriminate the two classes based on their pathological complete response (pCR) to drug X we will use all the metrics described in chapter 5.1.6. According to these metrics, an informative matrix like the one depicted in Table 5 and a graphical representation of the ROC curve as in Figure 6 are given to the researcher. For our classifier, a boxplot with mean values and standard errors showing the classification measures through the iterative cross-validation will be represented as well.

Metric	Mean	Standard Deviation
Accuracy		
Sensitivity		
Specificity		
Precision		
Recall		
AUC		
Kappa		

Table 5 Assessing the classification performance

The researcher either works directly to the INTEGRATE platform or downloads all the analysis to his/her local computer. The downloaded analysis could be an excel file with the resulted tables and graphical results placed in the same sheet.

5.3 Scenario B-Retrospective use of data

Scenario B	
Objective	A researcher wants to define if a gene expression signature can be used to predict the toxic effects (grade 3 (G3) or more) of an investigational class of drugs (e.g. mAb, TKI) used in the neoadjuvant treatment of a specific breast cancer subtype.
Steps	• The researcher logs into the system. • The researcher filters by type of cancer (i.e. breast), the treatment setting (i.e. neoadjuvant), the selected class of treatment and the toxicity (i.e. G3 or more). • The researcher selects for the following outputs; gene expression data, type of drug, toxicity type and grade, clinical trial and patients baseline characteristics. • The researcher either downloads the results on his computer (i.e. an excel file in csv format) and the gene expression data in the relevant format or works directly on the INTEGRATE platform using the provided tools.
Results	The researcher analyses gene expression data and tries to confirm his hypothesis: “A gene signature can predict the toxicity of a class of drugs”.

Table 6 Scenario B-Retrospective use of data

The objective of this study is slightly changed compared to the previous one, but the overall prediction framework remains the same. As in previous scenarios, the researcher logs into the INTEGRATE platform and exports the examined dataset which consists of patients with breast cancer, treated by an investigational class of drugs with a specific toxicity type and grade per drug, under neoadjuvant therapy. All available patients that received the investigational drug are dichotomised into two classes based on the toxicity grade of the drugs; a class with high grade (grade 3 or more) and a class with low grade toxic effect. The overall dataset enters the prediction model as in Figure 7 and the researcher can get access to the results by an excel file with all the available information as mentioned in the description of the previous scenario (chapter 5.2).

5.4 Scenario C-Retrospective use of data

In this scenario, data generated from samples collected in a neo adjuvant clinical trial run by Institute Jules Bordet (IJB), are used to create a model that predicts response/resistance to a specific preoperative drug (i.e. epirubicin) in estrogen receptor-negative (ER-) breast cancer patients. Again the dataset is dichotomized based on the independent clinical variable of pCR. The researcher logs into the INTEGRATE platform and exports the examined dataset which consists of patients with breast cancer, treated by a specific preoperative drug (i.e. epirubicin), under neoadjuvant therapy. This scenario encapsulates the adaptation of both kernel-based and ensemble trees classification techniques for prediction analysis.

When using kernel-based approaches, both implementations as depicted in Figure 4, can be applied to the examined dataset (a basis kernel for each data source or a basis kernel for each feature). Specifically, a first model of the MKL using support vector machines will compute separately a basis kernel for each data source (i.e. a kernel for each clinical, microarray and proteomic dataset respectively) and a unique kernel from

the linear combination of the individual kernels projects the overall data to the feature space for further analysis. A prediction model is then constructed by using all the available data (no feature analysis is performed), and its accuracy is assessed using the statistics in chapter 5.1.6 under the iterative validation techniques described in chapter 5.1.7.

A slightly changed kernel-based framework will be constructed using an individual kernel for each feature of all the available data sources. As in both scenarios described in 5.2 and 5.3, feature selection for reducing the high dimensionality of the microarray and molecular data will be first implemented. Then, a weighted basis kernel is computed for each feature and an iterative analysis is performed to estimate the most relevant features that give the highest classification accuracy.

A third prediction model will be provided using the ensemble of the decision trees (see chapter 5.1.5 for further details). Using several methods from the field of decision trees like the RotBoost, Random Forest and Rotation Forest, we aim in assessing the performance of them using an iterative evaluation process (i.e. bootstrapping or cross validation) and choose the “tree model” that shows the maximum performance.

Finally, the most accurate prediction model from both fields will be selected, becomes a part of the INTEGRATE platform and could be used as a predictive model for response to a specific preoperative drug (i.e. epirubicin) in estrogen receptor-negative (ER-) breast cancer patients.

6 Conclusion

This deliverable summarised the main objectives of the WP proposing an approach of building the framework for INTEGRATE VPH predictive modelling development. The clinically relevant questions that have been defined so far with concern the development of scenario-driven prediction models that given a set of characteristics will be able to predict in an accurate way the response to a drug and/or the response/resistance to a specific preoperative drug. This deliverable highlighted the main techniques that will be exploited giving emphasis to multi-kernel techniques that will allow the integration of multi-level heterogeneous data and subsequently the development of predictive models beyond the sate-of-the-art.

7 Appendix

Additional scenarios that do not represent highest priorities for users but they pose interesting challenges from the computational or methodological view, and will therefore by utilized for the definition of the generic framework, are presented in the following chapters. A statistical analysis of the heterogeneous data along with the implementation of the prediction modelling framework offers a thorough study to the researchers and constitutes a relevant and tool for assessing the clinical behaviour of patients' response to disease.

7.1 Scenario D-Retrospective use of clinical data

Scenario D	
Objective	An academic researcher wants to compare the pathological complete response (pCR) rate obtained using two different treatment regimens in the neo-adjuvant setting in a specific breast cancer subtype.
Steps	• The researcher enters the system. • The researcher filters the entire dataset by type of cancer (i.e. breast), the treatment setting (i.e. neoadjuvant) and the pathological characteristics (HER2+). • The researcher selects the treatment type and the pathologic response. • The researcher either downloads the results on his computer (i.e. an excel file in csv format) or works directly in the INTEGRATE platform using the provided tools and defines the rate of pCR in HER2+ patients treated with a standard regimen VS a regimen containing an investigational drug.
Results	The results will be used to design the statistical hypothesis for a new NeoBIG trial.
Example	A researcher wants to define the response to a standard regimen VS a regimen containing an investigational drug (monoclonal antibodies (mAb), tyrosine kinase inhibitors (TKI)) in human epidermal growth factor receptor 2 positive (HER2+) breast cancer patients using pCR as endpoint.

Table 7 Scenario D-Retrospective use of clinical data

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. The user logs into the INTEGRATE platform and the examined dataset is exported, using queries for filtering data by type of cancer, treatment settings and the pathological characteristics. The pCR, a binary value that corresponds to the disappearance or not of tumour at the tissue level evaluated at surgery, will be used as a criterion for the assessment of the regimen efficacy. A statistical analysis that enables the researcher to quantitatively specify whether or not the investigational regimen leads to better results will be done under a web-based unified framework or by executable programming logic.

From the technical aspect, the pCR rate as well as the Odds Ratio though forest plots [50] will be employed to measure the investigational regimen effect versus standard regimens effect. In order to estimate the pCR rate, the number of patients with pCR is divided by the total patients who received a standard and investigational regimen,

respectively. The pCR rate will be a basic approach for estimating the percentage of pCR in a subgroup of patients who received a specific type of regimen. The odds ratio will be given by considering a trial in which a number of patients were randomized to two different regimens, as depicted in the following table.

	Standard Regimen	Investigational Regimen
pCR	a	b
No pCR	c	d

Table 8 2x2 table for odds ratio estimation

This study will be extended towards the assessment of strength dependence between data values within each subpopulation that received a specific type of regimen. The available clinical data, presented in chapter 4.1, will contribute to the calculation of the following characteristics, as seen below.

Characteristic	No. of Patients	Patients with pCR (%)
Age, years		
≤50	86	10.5
>50	53	18.9
Tumor Size		
T1-T2	119	13.4
T3-T4	20	15.0
Ki67		
≤25	23	8.7
>25	92	15.2

Table 9 Clinical Characteristics for Evaluable Patients treated with Anthracyclines

Given the characteristics above, one can estimate the strength of dependence between the characteristics and the pCR (i.e. significant association is found between patients with tumor size T1-T2, achieving pCR when treated with anthracyclines). A statistical analysis can therefore be provided by the INTEGRATE platform, in which odds ratios are estimated between the clinical characteristics and the pCR of patients treated with a standard and an investigational regimen, respectively. Conclusively, the platform returns a ranked table with the clinical-pCR dependence for every regimen having a format as the presented table below.

	Standard Regimen			Investigational Regimen		
Characteristic	**OR**	**95% CI**	**p value**	**OR**	**95% CI**	**p value**
Age, years						
≤50						
>50						
Tumor Size						
T1-T2						
T3-T4						
Ki67						
≤25						
>25						

Table 10 Representation of odds ratios for both regimens

The INTEGRATE platform will also provide a graphical representation of the odds ratio between the characteristics and the pCR, for every type of regimen. An example is given in the following figure, represented from [51].

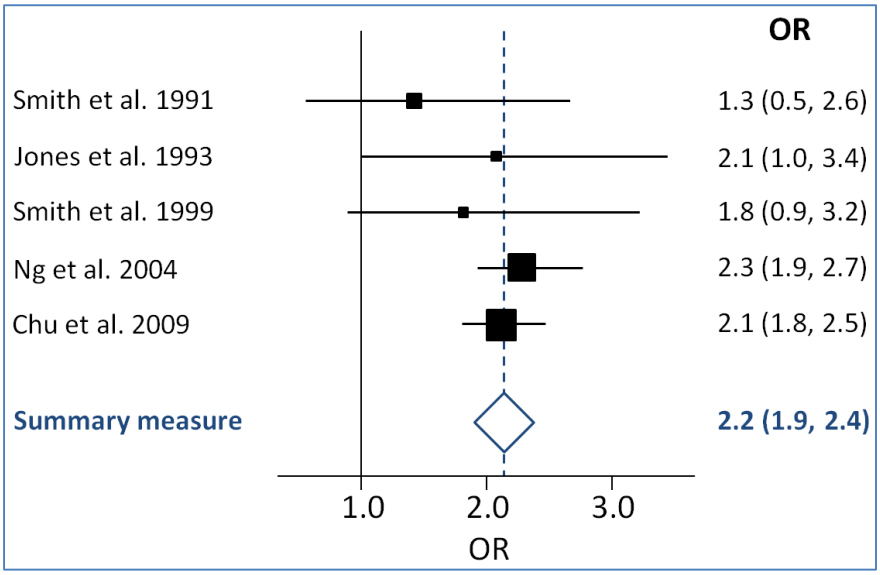

Figure 8 Forest plot of odds ratios and associated confidence intervals [51]

The overall statistical analysis (pCR rate and Odds ratios with CIs and p-values) will be available for download on a local computer (i.e. an excel file in csv format). All p-values will be two-sided with statistical significance evaluated at the 0.05 alpha level and confidence interval of 95% will be calculated to assess the precision of the obtained estimates.

7.2 Scenario E-Retrospective use of clinical data

Scenario E	
Objective	An academic researcher wants to define if pCR is a candidate surrogate marker for Disease Free Survival (DFS) and Overall Survival (OS) independently of treatment type.
Steps	• The researcher logs into the system. • The researcher filters by type of cancer (i.e. breast), the treatment setting (i.e. neoadjuvant), selected treatment (i.e. all) and the pathological characteristics (i.e. all). • The researcher selects the outcome data (i.e. pCR, DFS, OS). • The researcher either downloads the results on his computer (i.e. an excel file in csv format) or works directly in the INTEGRATE platform using the provided tools and defines how pCR correlates to DFS and OS independently of treatment type.
Results	According to the obtained results, the academic researcher will design a new NeoBIG trial in which pCR will or will not be used as a surrogate endpoint.

Table 11 Scenario E-Retrospective use of clinical data

In this scenario, a researcher wants to investigate the association between the pathological complete response (pCR) and clinical outcome in terms of Disease Free

Survival (DFS) and Overall Survival (OS) independently of any treatment type. According to the obtained results, the researcher will design a new NeoBIG trial in which pCR will or will not be used as a surrogate endpoint.

Initially, the researcher logs into the INTEGRATE platform and exports the examined dataset which consists of patients with breast cancer, treated by any possible type of regimen under neoadjuvant therapy. The examined dataset is dichotomized into two groups using the binary variable pCR as the independent variable. In complex diseases, such as cancer, researchers rely on statistical comparisons of DFS and OS of patients against healthy control groups or against patients following different treatment as in [52]. In this approach DFS and OS will be estimated by Kaplan-Meier survival analysis [53] and the log-rank test will be used to compare DFS and OS between the two groups (pCR VS no pCR achieved).

In this statistical analysis, Kaplan-Meier survival curves, along with the 95% confidence interval for the curves, showing the DFS and OS of pCR and non-pCR patients treated by any type of neoadjuvant regimen will be presented. A representative example, given by [54], is illustrated below. Anthracycline-treated patients were dichotomized by their Tissue Inhibitor of Metalloproteinases-1 (TIMP-1) level and Kaplan-Meier survival analysis has been done, showing the cumulative percentages of DFS (subfigure A) and OS (subfigure B) over time.

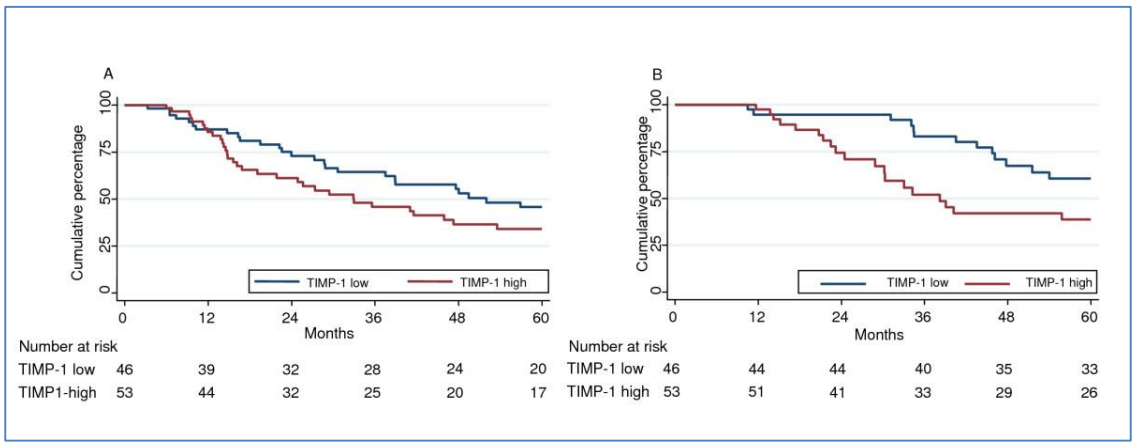

Figure 9 Kaplan-Meier plot showing the DFS (A) and OS (B) of TIMP-1 [54].

To compare the survival distributions given by our groups, the widely-used non-parametric log-rank test will be performed. It provides a p value that indicates whether or not the difference in survival between the two groups is statistically significant. Therefore, estimating the log-rank between the survival curve of pCR and non-pCR groups we interpret a p value that indicates a statistically significant deference (low p values) or a convergence of the two curves if the p value is high.

After the completion of the statistical analysis, the researcher either works directly to the INTEGRATE platform or downloads all the analysis to his/her local computer. The downloaded analysis could be an excel file with the resulted tables and graphical results placed in the same sheet.

7.3 Scenario F-Retrospective use of imaging data

Scenario F	
Objective	An academic researcher wants to define if pCR is associated with a decrease of more than 20% in tumor volume between baseline and day 15, so that decrease in tumor volume between baseline and day 15 could be used as an early surrogate for pCR.
Steps	• The researcher logs into the system. • The researcher filters by type of cancer (i.e. breast) and the treatment setting (i.e. neoadjuvant). • The researcher selects for these outputs: the imaging data at baseline and at day 15 and response to treatment. • The researcher either downloads the results on his computer (i.e. an excel file in csv format) or works directly in the INTEGRATE platform using the provided tools and defines if pCR is associated with a decrease of more than 20% in tumor volume between baseline and day 15 or not.
Results	According to the obtained results, the researcher will design a new NeoBIG trial to validate if the decrease in tumor volume between baseline and day 15 could be used as an early surrogate marker for pCR.

Table 12 Scenario F-Retrospective use of imaging data

Tumor volume can be extracted from the information contained in the tags of standard DICOM images (MRI, CT etc.) and the delineation/segmentation of the tumor by the doctors, using the "DrEye" tool by FORTH [55]. DICOM stands for Digital Imaging and Communications in Medicine, and it is a standard for handling, storing, printing, and transmitting information in medical imaging. It includes a file format definition and a network communications protocol. The National Electrical Manufacturers Association (NEMA) holds the copyright to this standard. It was developed by the DICOM Standards Committee, whose members are also partly members of NEMA.

The DICOM format groups information into data sets. That means that a file of a chest X-Ray image, for example, actually contains the patient ID within the file, so that the image can never be separated from this information by mistake. This is similar to the way that image formats such as JPEG can also have embedded tags to identify and otherwise describe the image. A DICOM data object consists of a number of attributes, including items such as name, ID, etc., and also one special attribute containing the image pixel data (i.e. logically, the main object has no "header" as such: merely a list of attributes, including the pixel data). A single DICOM object can only contain one attribute containing pixel data. For many modalities, this corresponds to a single image. But note that the attribute may contain multiple "frames", allowing storage of cine loops or other multi-frame data. Another example is NM data, where an NM image by definition is a multi-dimensional multi-frame image. In these cases three- or four-dimensional data can be encapsulated in a single DICOM object. Pixel data can be compressed using a variety of standards, including JPEG, JPEG Lossless, JPEG 2000, and Run-length encoding (RLE). LZW (zip) compression

can be used for the whole data set (not just the pixel data) but this is rarely implemented.

The estimation of the tumor's volume is based on the fact that the tumor is a collection of voxels. A voxel can be defined as the volume unit, which can be computed from the information contained in the DICOM tags of each image in the series of interest. The volume of a voxel is the product of the information in the tag $(0028,0030)$ by the sum of the information in the tag $(0018,0050)$ and in tag $(0018,0088)$. These calculations are given by the following equation:

$$Volume = PixelSpacing.Width * PixelSpacing.Height * (SliceThickness + SpacingBetweenSlices)$$

With the volume unit (=voxel) to be defined, a draft estimation of volume of the tumor is computed based by multiplying the sum of all the voxels of the tumor with the volume unit. The estimation of the volume can be improved by interpolating the available series and creating a new series in an isotropic space where the volume unit is defined as 1 cubic mm, which is fairly smaller than the one without interpolation.

Having the available information from the DICOM imaging system, we estimate the tumor volume at baseline and day 15 for each patient. All cases are dichotomised based on if they achieve to perform a decrease in volume change equal or more than 20%. A confusion matrix is then given between the pCR and the decrease of tumor change, as depicted in Table 13. An odds ratio statistical analysis will be performed to characterize the association between the two variables and validate if the decrease in tumor volume between baseline and day 15 could be used as an early surrogate marker for pCR. A graphical representation of the odds ratios through forest plots will be also provided, as depicted in Figure 8.

	≥20% decrease	<20% decrease
pCR	a	b
No pCR	c	d

Table 13 Confusion matrix for tumor volume change

8 REFERENCES

1. P. Carmeliet, R. K. Jain, “Angiogenesis in cancer and other diseases”, Nature, vol. 407 pp. 249-259, 2002.
2. Cristini, V., Frieboes, H.B., Gatenby, R., Caserta, S., Ferrari, M., Sinek, J.P. Morphological instability and cancer invasion. Clin. Cancer Res. 11, 6772–6779, 2005.
3. http://en.wikipedia.org/wiki/Virtual_Physiological_Human
4. Fenner JW, Brook B, Clapworthy G, Coveney PV, Feipel V, Gregersen H, Hose DR, Kohl P, Lawford P, McCormack KM, Pinney D, Thomas SR, Van Sint Jan S, Waters S, Viceconti M. The EuroPhysiome, STEP and a roadmap for the virtual physiological human. Philos Transact A Math Phys Eng Sci. 366(1878):2979-99, 2008.
5. L. B. Edelman, J. A. Eddy, and N. D. Price, In silico models of cancer. Wiley & Sons, 2009.
6. Dedeurwaerder S, Desmedt C, Calonne E, Singha SK, Haibe-Kains B, Defrance M, Michiels S, Volkmar M, Deplus R, Luciani J, Lallemand F, Larsimont D, Toussaint J, Haussy S, Rothé F, Rouas G, Metzger O, Majjaj S, Saini K, Putmans P, Hames G, Baren NV, Coulie PG, Piccart M, Sotiriou C, Fuks F. DNA methylation profiling reveals a predominant immune component in breast cancers. EMBO Mol Med. 2011.
7. Uwe Scherf et al., A gene expression database for the molecular pharmacology of cancer, nature genetics, volume 24, March 2000.
8. G. Stamatakos, D. Dionysiou, N. Mouravliansky, K. Nikita, G. Pissakas, P. Georgolopoulou and N. Uuznoglu, “Algorithmic Description of the Biological Activity of a Solid Tumor in Vivo”, in Proc. EUROSIM 2001 Congress, Delft, the Netherlands, June 26-29, 2001 (CD-ROM Edition).
9. K.Swanson, E.C.Alvord Jr., J.D.Murray, “Dynamics of a model for brain tumors reveals a small window for the therapeutic intervention,” Discrete and Continuous Dynamical Systems-Series B, vol. 4, no 1, pp.289-295, 2004.
10. Olivier Glatz, Maxime Sermesant, Pierre-Yves Bondiau, Hervé Delingette, Simon K. Warfield, Grégoire Malandain, and Nicholas Ayache. Realistic Simulation of the 3D Growth of Brain Tumors in MR Images Coupling Diffusion with Mass Effect. IEEE Transact. on Medical Imaging, 24(10):1334-1346, 2005.
11. Benjamin Ribba, Thierry Colin and Santiago Schnell, A multiscale mathematical model of cancer, and its use in analyzing irradiation therapies, Theoretical Biology and Medical Modelling, 3:7, 2006.
12. Graham A. Colditz and A. Lindsay Frazier, Models of Breast Cancer Show That Risk Is Set by Events of Early Life: Prevention Efforts Must Shift Focus, Cancer Epidemiology, Biomarkers & Prevention, Vol. 4. 567-571. July/August 1995.
13. Schmid P, Wischnewsky MB, Sezer O, Böhm R, Possinger K:Prediction of Response to Hormonal Treatment in Metastatic Breast Cancer. Oncology; 63:309-316, 2002.
14. Wolmark N, Wang J, Mamounas E, Bryant J, Fisher B. Preoperative chemotherapy in patients with operable breast cancer: nine-year results from National Surgical Adjuvant Breast and Bowel Project B-18. Journal of the National Cancer Institute, (30):96-102, 2001.
15. C Desmedt, A Di Leo, E de Azambuja, D Larsimont, B Haibe-Kains, J Selleslags, S Delaloge, C Duhem, J-P Kains, B Carly, M Maerevoet, A Vindevoghel, G Rouas, F Lallemand, V Durbecq, F Cardoso, R Salgado, R

Rovere, G Bontempi, S Michiels, M Buyse, J-M Nogaret, Y Qi, F Symmans, L Pusztai, V D'Hondt, M Piccart-Gebhart and C Sotiriou. Multifactorial Approach to Predicting Resistance to Anthracyclines. American Society of Clinical Oncology, Volume 29, Number 12, April 2011.

16. 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.
17. 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.
18. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507-2517, 2007.
19. Yee Hwa Yang, Yuanyuan Xiao, and Mark R. Segal. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics* 21, 1084-1093, April 2005.
20. S. Yu, T. Falck, A. Daemen, L.-C. C. Tranchevent, J. A. A. Suykens, B. De Moor, and Y. Moreau. L2-norm multiple kernel learning and its application to biomedical data fusion. BMC bioinformatics, vol. 11, no. 1, pp. 309+, Jun. 2010.
21. Wilcoxon, Frank. Individual comparisons by ranking methods. Biometrics Bulletin 1 (6): 80–83, Dec. 1945.
22. J. W. Lee, J. B. Lee, M. Park, and S. H. Song. An extensive comparison of recent classification tools applied to microarray data. Comp. Statistics and Data Analysis, vol. 48, pp. 869–885, 2005.
23. Sutton, C.D. Classification and Regression Trees, Bagging, and Boosting. Handbook of Statistics 24, 303–329, 2005.
24. Gevaert, O. A Bayesian network integration framework for modeling biomedical data. Ph.D dissertation, Katholieke Universiteit Leuven, 2008.
25. Vapnik, V. The Nature of Statistical Learning Theory. Springer, N.Y. 1995.
26. B. Scholkopf, A. J. Smola. Learning with Kernels. The MIT Press, 2002.
27. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Smola, A.J., Mueller, K.-R. Constructing descriptive and discriminative non-linear features: Rayleigh coefficients in kernel feature spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004.
28. J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
29. 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.
30. P. Pavlidis, J. Cai, J. Weston and W. Noble Grundy. Gene functional classification from heterogeneous data. In Proceedings of the Fifth Annual International Conference on Computational Biology: April 22-25, 2001**.**
31. Pai-Hsuen Chen, Chih-Jen Lin and Bernhard Scholkopf, “A Tutorial on v-Support Vector Machines”. Applied Stochastic Models in Business and Industry, Vol. 21, pp. 111-136, No. 2. 2005.
32. A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans. Advances in Large Margin Classifiers. MIT Press, Cambridge, MA, 2000.
33. Breiman, L., Bagging predictors. Machine Learning 24 (2), 123–140, 1996.
34. Freund, Y., Schapire, R.E., Experiments with a new boosting algorithm. In Proceedings 13th International Conference on Machine Learning. Morgan Kaufmann, Bari, Italy, pp. 148–156, 1996.
35. Breiman, L., Random forests. Machine Learning 45 (1), 5–32, 2001.

36. Rodríguez, J.J., Kuncheva, L.I., Alonso, C.J. Rotation forest: A new classifier ensemble method. IEEE Trans. Pattern Anal. Machine Intell. 28 (10), 1619–1630, 2006.
37. Meir, R., Rätsch, G. An introduction to boosting and leveraging. In: Advanced Lectures on Machine Learning. Lecture Notes Comput. Sci., vol. 2600. Springer-Verlag, Berlin, pp. 118–183, 2003.
38. Jin, R., Zhang, J. Multi-class learning by smoothed boosting. Machine Learning. 67 (3), 207–227, 2007.
39. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
40. Smeeton, N.C. Early History of the Kappa Statistic. *Biometrics* **41**: 795, 1985.
41. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 33:159–174, 1977.
42. David, A., Comparison of classification accuracy using Cohen's weighted kappa. Expert Syst. Appl. 825-832, 2008.
43. Fawcett, T. An introduction to ROC analysis, Pattern Recognition Letters 27(8),861-874, 2006.
44. Z. Chen, J. Li, and L. Wei. A multiple kernel support vector machine scheme for feature extraction and rule extraction from gene expression data of cancer tissue. Artificial Intelligence in Medicine, vol. 41, no. 2, pp. 161-175, October 2007.
45. M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. Proceedings of the IEEE International Conference on Computer Vision, pp. 1-8, October 2007.
46. A. Zien and C. S. Ong. Multiclass multiple kernel learning. Proceedings of the 24th International Conference on Machine Learning, pp. 1191-1198, October 2007.
47. Ye, J., Alexander, G., Wu, T., Chen, K., Wu, T., Li, J., Zhao, Z. Heterogeneous data fusion for Alzheimer's disease study. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 08, 2008.
48. Ran Gilad-Bachrach, Amir Navot and Naftali Tishby. Margin Based Feature Selection-Theory and Algorithms. Proceedings of the 21st International Conference on Machine Learning, Canada, p. 43-50, 2004.
49. K.Kira and L.Rendell. A practical approach to feature selection. Proceedings of 9th International Workshop on Machine Learning, p 249-256, 1992.
50. S. Lewis and M. Clarke. Forest plots: trying to see the wood and the trees. BMJ (Clinical research ed.), 322(7300):1479–1480, June 2001.
51. http://en.wikipedia.org/wiki/Forest_plot
52. S. Kaura and G. Dranitsaris. Letrozole or anastrozole for the treatment of hormone-positive breast cancer: A clinical comparison using indirect statistical techniques. American Society of Clinical Oncology, Volume 28, Number 15, May 2010.
53. Kaplan, E.L., Meier, P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481 1958.
54. Schrohl AS, Look MP, Meijer-van Gelder ME, Foekens JA, Brünner N. Tumor tissue levels of Tissue Inhibitor of Metalloproteinases-1 (TIMP-1) and outcome following adjuvant chemotherapy in premenopausal lymph node-positive breast cancer patients: A retrospective study. BMC Cancer. 10; 9:322, Sep. 2009.
55. http://biomodeling.ics.forth.gr