Model Driven Paediatric European Digital Repository

Call identifier: FP7-ICT-2011-9 - **Grant agreement no**: 600932

Thematic Priority: ICT - ICT-2011.5.2: Virtual Physiological Human

Deliverable 12.1

Outline of the clinical assessment and validation criteria for all four disease areas

Due date of delivery: 31st August 2014

Actual submission date: 30th October 2014

Start of the project: 1st March 2013
Ending Date: 28th February 2017

Partner responsible for this deliverable: OPBG

Version: 1.4

Dissemination Level: Public

Document Classification

Title	Outline of the clinical assessment and validation criteria for all four disease areas
Deliverable	D12.1
Reporting Period	Period 2
Authors	OPBG
Work Package	12
Security	PU
Nature	RE
Keyword(s)	Model validation, clinical assessment

Document History

Name	**Remark**	**Version**	**Date**
OPBG	Preliminary Version	V1.1	27th August 2014
LYNKEUS	Revised Draft	V1.2	10th September 2014
OPBG	Draft sent out to the relevant partners	V1.3	20th September 2014
LYNKEUS	Includes all the comments, feedback and contributions by the relevant partners	V1.4	28th October 2014

List of Contributors

Name	**Affiliation**
Marcello Chinali	OPBG
Alex Jones	UCL
Jakob Hauser	UCL
Alexey Tsymbal	SAG
Pieter van Dijkhuizen	IGG
Joe Prinold	USFD
Jaap Harlaar	VUmc
Marjolein van der Krogt	Vumc
Patrick Ruch	HES-SO
Omiros Metaxas	ATHENA RC
Harry Dimitropoulos	ATHENA RC
Maxime Semersant	INRIA
Stefan Wesarg	Fraunhofer
David Manset	GNUBILA

List of reviewers

Name	**Affiliation**
Frans Steenbrink	MOTEK
Marco Viceconti	USFD
Giacomo Pongiglione	OPBG
Edwin Morley-Fletcher	LYNKEUS
Bruno Dallapiccola	OPBG

Table of Contents

1. Aim of the present document

The aim of the Work Package 12 is to assess the accuracy of multidisciplinary derived models in order to improve the identification of markers of outcome prediction and risk stratification, and thus to derive and evaluate personalized treatment models.

The goal is to validate the computational models to assure that they can be personalised by adapting the parameters to the integrated data of a specific patient and to improve the current knowledge and understanding of the disease by simulating different aspects on the evolution of a disease. In addition it also aims at verifying the accuracy of the insights of the effect of a specific therapeutic intervention; being this either pharmacological, behavioral or surgical.

Clinical validation of the models, as stated in the description of work, will be an ongoing process which benefits from the use of the models in different clinical settings thus increasing the number of clinical observations and improving the stability of the clinically derived model. The final aim of WP12 is to assure that the data repository will be in the condition to be continuously improved, in order to assure the accuracy and stability of the derived models and result into integrated clinical workflows leading to personalized treatment models.

The aim of the present document is to define an outline of the clinical validation criteria process for the four disease areas. From the preliminary stage of validation methodology-outline, described in detail below, validation reports performed during the study progression will contribute to define eventually model-driven clinical workflow, through a continuous evolving process which will not only bring updates to the validation process but also represent elements of its refinement.

In agreement with the approach outlined in the Quality Assurance Guidelines (D1.4), this document will take into account the diversity of the participating project partners. Accordingly, the validation process does not seek to override existing clinical and technical procedures, with specific regard to models validation methodologies developed by the relevant technological partners. Thus, the validation criteria and process presented in this document will be flanked by the internal and partner-specific validation practices.

Finally, in order to ensure the maximum level of final acceptability and usability of these models within the clinical community, the Users Board will be actively involved in the validation process.

2. General principles of clinical assessment and validation

As stated in D19.2 (HTA evaluation framework), before the application and deployment of a new technology in a real context, it is necessary to proceed to an assessment of the Technology Readiness Level (TRL), which can be appropriately evaluated only through subsequent experimentation, refinement, and increasingly realistic clinical testing and validation.

The clinical validation is the assessment and analysis of the accuracy of the derived model to accurately predict the outcome of the patients. Thus from a clinical perspective, it assures that the computer and mathematical models derived from the single four disease areas achieve their intended performance.

To conduct the clinical assessment, a number of different aspects will be taken into account:

1) *Identify Clinical data relevant to the models to assure they answer the clinical questions;*
2) *Identify additional clinical data, included in the data collection clinical protocols, that are not included in building the models, but that during the validation process will be deemed as useful to improve the models and thus used to address potentially unexpected unsolved issues;*
3) *Bring all the clinical data together to reach conclusions about the performance of the models*.

3. General outline of the Clinical Assessment and Validation

3.1 Introduction and general principles

The clinical assessment and validation will be based on a comprehensive analysis of available clinical data for the four disease areas, clinical performance data, and safety data. The appropriate validation of the models obviously depends on the purpose of its intended use. Having four different disease areas, significant heterogeneity in the building and aim of the different models exist.

Validations for all disease areas will provide assurance that predictions made by the models are correct. The computer model internal validation will include both retrospective and prospective validations. Validations will be subdivided into internal validation and external validation, as detailed below. Furthermore, the validation will follow a multi-layer approach: Partial Validation, Full-scale Validation and Cross-validation.

- *The first level of validation will be the initial testing and debugging, which will be performed in close collaboration with the technical partners.*

- *The second level will be the internal validation (both prospective and retrospective) that will ensure that the different mechanistic models reproduce results of the clinical studies used to build the single model for the specific disease area.*
- *The third level of validation will be the external validation which will define the ability of the model to accurately predict the results of studies acquired through the digital repository and thus to derive a statistical model from data that were not used to build the model.*

Internal validations will apply, as precisely as possible, to the study population and the protocol of the clinical WP that generated the data for validation. Validations will report the absolute value of the predicted outcome variables and the observed difference in outcomes between groups (modelled vs observed). Each validation will involve the complete spectrum of subjects in each single disease area. Ideally, we aim at obtaining and use patient-specific data for the single disease areas.

Evaluation of the clinical validation will take into account the effects of sampling error, which reflects the sample size of the single WP (i.e. the validation) cohort. Confidence intervals (CIs) around the observed results in the validation cohorts will be reported. Both tabular and graphical reporting of validation analyses will be performed, to provide both statistical and visual reports of validation process of the single models.

Models will be externally validated against as many clinical studies as possible, through the digital repository. The digital repository is a first layer of what is expected to become a generic information technology framework for the collection, organization, and utilization of heterogeneous medical information ranging from sources such as Electronic Medical Records (EMR). The aim of the repository is to provide the creation and maintenance of Patient-Specific Models, providing a multi-scale, comprehensive, precise, personalized representation of the patient and to provide real-time decision support system, promoting optimized diagnostic, prognostic and therapeutic decisions throughout the treatment workflow.

Multiple external validations will be needed to significantly improve the effectiveness and increase the stability of the derived models. This is because we expect that, although very accurate in the internal validation, models are possibly going to fail in some individual patients, either because of unanticipated features of the individual patient or because of the counfounding effect in some clinical settings caused by local differences in treatment or general clinical approach. It is also worth to notice that in some instances the models might occasionally succeed simply by chance. Thus, a large number of successes and near misses derived from a large number of observations is expected to represent a good indicator of the fundamental accuracy of the model. Validating a model against a large number of studies through the digital repository also excludes the hypothesis that models are not only working on specific studies used in the model due to specific selection biases.

If validation against a particular study fails, efforts will be made to determine why the failure occurred and consider revising the model to achieve a match. If this is accomplished, then the exercise will not be considered as successful external validation but defined as a recalibration. When a model is recalibrated or redesigned to fit a specific subset of studies, steps will be taken to ensure that the revised model is still valid for all the previous validation exercises against other clinical studies.

Since we believe that the initially derived models will not be successful at every attempt at an external validation, successive attempts though calibration and retesting are still expected to yield a series of successes and failures, leading to improved versions of the model and to gain that the proportion of successful external validations will significatly increase.

In summary, the clinical assessment of the model will result in:

a) Maximal accuracy of the model

b) Identification of strongest markers of outcome prediction and

c) Insights into personalized treatment models and model driven clinical workflows.

3.2 Initial Testing and Debugging

With a view on user validation and interfaces quality improvement, e-infrastructure applications supporting MD-Paedigree clinical models, released by the Infostructure follow a Scrum agile process whereby user representatives are consulted on the products features, improvements and bug fixes to be dealt with. Scrum is an iterative and incremental agile software development framework for managing products developments. It defines "a flexible, holistic product development strategy where a development team works as a unit to reach a common goal", challenges assumptions of the "traditional, sequential approach" to product development, and enables teams to self-organize by encouraging co-location or close online collaboration of all team members, as well as frequent face-to-face communication among all team members and disciplines in the project. This way, the foundational Infostructure provides a solid technical base onto which new clinical models can be developed and tested.

Thus, clinical assessment and validation of the project models, starts with internal testing and debugging as the first performed analysis. It will indeed analyse the models built from the clinical WPs and verify the clinical accuracy of the predicted outcome.

Since MD-Paedigree is a clinically-driven project, initial testing will assure that the models comply with the clinical user requirements and that the information obtained from the models provide clinically useful information and predictions.

In order to perfom the initial testing, models will be tested in a number of real and hypothetic clinical settings in which the outcome is known or can be easily foreseen in clinical practice. For example in a patient with dilated cardiomyopathy in which, despite aggressive medical therapy, a chronic and severe decline in functional parameters of heart function in observed, associated with increasing severity of heart failure symprtoms, it is expected that progression of the disease will lead to need of mechanical heart support, heart transplant or exhitus within one year.

The aim of the initial testing is to assure that the models are able to provide information on both the evolution and the outcome of the single diseases.

At this stage there will be no statistical assessment of the accuracy of the models but rather a general verification performed by the different disease area clinical experts in order to establish whether the derived models are providing information on the specific test patients that are in agreement with known and scientifically established pathophysiology patterns for the single diasese areas, that the information provided are useful from a clinical standpoint, and that the information provided add to current knowledge of the specific disease areas.

3.3 Internal Validation

The second stage of the clinical assessment and validation will be the internal validation. To perform the validation and verification, a list of preliminary aspects of the models will be taken into account:

- Data selectivity/specificity
- Accuracy and precision
- Repeatability and Reproducibility
- Limit of detection and of quantification

3.3.1 Data selectivity and specificity

The first step of the internal validation will be to verify the data selectivity and specificity of the models. The ability of the models to deliver information on the development and prognosis of the single diseases as well as to predict the effect of treatment strategy, relies on the ability of the different models in identifying relevant variables and relevant interaction among variables (i.e. specificity) and discriminate from confounding data available (i.e. selectivity). Thus model specificity/selectivity refers to the extent to which it can determine key factors in a complex mixture of data without interference from other components. Being that MD-Paedigree is a clinically-led study, data will be collected in each disease areas according to ‘real life’ clinical workflow, and thus will include a number of clinical, anthropometric, laboratory and diagnostic

imaging data, that do not have all the same impact on disease progression and prognosis. Accordingly, the model will be tested in order to verify its ability to identify key components of the disease and relevant interactions among variables, disregarding the effect of noise and confounding effect of the multitude of clinical and diagnostic data collected. Thus, the models will be tested on their ability to provide reproducible predictions of the single disease by taking into account principal components of the disease despite the confounding effect of minor components. To establish selectivity and specificity of the models we have preliminary identified - by means of the user requirement analysis reported in D2.1 - a number of key endpoint variables for each disease. These variables include both primary endpoints (e.g. in WP3 hard endpoints have been defined as patient exhitus or need of heart transplant) and secondary endpoints. Secondary endpoints include key pathophysiological interactions which characterize different aspect of the single diseases, and that are derived for diagnostic techniques used in the different WPs (for example the effect of total amount of body fat content on vascular reactivity for WP4). Detailed list of data that will be tested for selectivity and specificity are provided in the single WP validation analysis outline below.

3.3.2 Accuracy, Precision, Reproducibility and Repeatability

Another aspect tested in the internal validation will be to establish the accuracy and precision of the models. The precision of the models will be related to their reproducibility and repeatability, in order to verify that repeated measurements under unchanged conditions provide the same results. Differences in repeated measurements will be tested by test-retest analysis and rho values reported. Once the accuracy of the model is established, the prediction of both primary and secondary endpoints will be tested for precision, to identify the possible effect of systematic errors in the prediction. We expect that at a preliminary stage of validation high precision might not be achieved, given the limited number of observations included in the single disease areas. This will possibly be the case for those WPs in which heterogeneity in the disease recruitment is present (e.g. WP6), while in homogenous subset of patients (e.g. WP3) precision is expected to be high already at this stage. When accuracy and precision will not be achieved, analysis of systematic error will be performed to achieve accuracy and, through the use of the digital repository, increasing sample size will be used to increase precision. We expect that through the elimination of possible systematic error and by means of increased sample size the models will achieve precision and accuracy in all disease areas.

3.3.3 Limits of detection and quantification

Limits of detection (LOD) and of quantification (LOQ) will also be defined for the single disease models. LOD and LOQ will be quantified in order to establish the lowest amount of useful data that can be used by the models to provide an accurate and precise prediction of single disease progression. Being the clinical data acquired from the single WP derived from clinically established workflows, it is expected that not all data will be available for every patient in each disease. Thus, it is expected that heterogeneity in the amount of data acquired for each patient will be found throughout the different disease areas. For example

in WP3, assessment of exercise tolerance test is a key factor in quantification of symptoms related to the cardiomyopathy disease and thus provides useful information on the impact of variables obtained from imaging techniques on exercise tolerance and prognosis. However it is expected that these data will not be obtained in all younger patients, given the lack of compliance related to the low age of part of the participants. Thus, in these patients the model needs to provide features and predict prognosis of the patients even without complete datasets. Thus analysis of LOD will be used to establish ability of the model to provide information with the lowest amount of data, while the LOQ will be defined as the lowest amount of data needed to provide model prediction with acceptable precision and accuracy.

4. Outline of the Statistical Approach for Validation

Internal validation will compare data derived from the model simulation results to those of the clinical study. Statistical methods for comparisons will be based on three different statistical approaches as outlined in brief below:

1) Hypothesis testing
2) Confidence intervals of the means.
3) Comparison on standardized distance in variance.

4.1 Hypothesis testing

The statistical hypothesis test is a standard method of statistical inference. Statistical hypothesis test will be used to determine whether the outcomes provided from the single disease area models lead to a rejection of the null hypothesis. This will help to establish whether outcome prediction on the evolution of the different diseases contain enough information to cast doubt on conventional knowledge, given that conventional wisdom has been used to establish the null hypothesis.

4.2 Comparison intervals of the means

The comparison of the intervals of means is a standard statistical approach to define similarity of significant differences between two samples of observations (in this case predicted form the computer model vs observed in clinical practice). In fact the confidence interval for the difference between two means specifies a range of values within which the difference between the means of the two populations may lie. The confidence interval for the difference between two means contains all the values of (μ_1 - μ_2) (the difference between the two samples means) which would not be rejected in the two-sided hypothesis test of H0: $\mu_1 = \mu_2$ against Ha: $\mu_1 \neq \mu_2$, i.e.H0: $\mu_1 - \mu_2 = 0$ against Ha: $\mu_1 - \mu_2 \neq 0$. If the confidence interval includes 0 we can say that there is no significant difference between the means of the two populations, at a given level of confidence.

4.3 Comparison on standardized differences

Another approach will be to compare standardised differences among the outcome predicted from the computer models and the ones derived from clinical observations, as previously suggested (Abbasemail, Rovira and Casanovas. "Validation by simulation of a clinical trial model using the standardized mean and variance criteria", 2005). For any outcome variable a standardized distance in mean (SDM) will be derived and the standardized distance in variance (SDS) will be calculated as:

$$-10\% < \mathrm{SDM} = \frac{\bar{x}_o - \bar{x}_s}{s_o} < 10\%,$$
$$-10\% < \mathrm{SDS} = \frac{S_o - S_s}{S_o} < 10\%,$$

where $\bar{x}_o$ is the observed mean value; *So* corresponds to observed standard deviation; $\bar{x}_s$ is the model derived mean value, and *Ss* is the model standard deviation of the *outcome variables*. The first validation criterion is that the standardized distance in standard deviation is 1SD minus or plus, as the formula indicates. The second criterion is that the standardized distance in variances is 10% minus or plus. The third criterion is that the model has to reproduce these results simultaneously for all outcome variables. Selection of the best model will be based on these sequentially applied criteria: (1) the model needs to meet the three validation criteria; and (2) the most parsimonious model (i.e., the model containing the lowest number of parameters) fulfilling the previous criteria would be considered as most efficient.

5. External Validation

Before using a computer derived prediction model, it is essential that its performance is tested and validated in patients that were not used to develop the model.

External validation will be performed on the models that will be considered the most efficient and accurate after the initial evaluation and the internal validation. External validation will be used to a) significantly improve the number of observations and thus to verify the accuracy of the models; b) the stability of the models; and c) the clinical usefulness of the derived models. External validation will be performed through the use of the Digital Repository by means of the 'Similarity-Search Based Decision Support System', that was developed in the Health-e-Child project (HeC Case Reasoner, [Manset et al., 2009]), for decision support in three domains: cardiology, neuro-oncology, and rheumatology and that is currently being extended with the MD-Paedigree data collection of WP4, to decision support in the domain of modelling cardiovascular risk in obese children and adolescents. During the MD-Paedigree study data collection, the Digital Repository is being populated with data provided by the different disease areas. However, the repository for paediatric cardiac diseases (i.e. PCDR) was developed and exploited by OPBG for paediatric cardiology before the MD-Paedigree data collection process and is already populated with thousands of data from different cardiac

paediatric diseases. Thus external validation for WP3 models will start earlier as compared to the other disease areas.

Similarity search techniques will be included in the Digital Repository, which map specific medical cases to pertinent patient/disease profiles. These profiles are used to adapt and optimize individual simulation models by transformations, as well as to explore their combinations and re-use in different disease areas. As stated in the DoW, a holistic scheme for model-driven personalized medicine will be developed to allow analysing and testing scientific hypotheses, predicting disease evolution and treatment responses (e.g. early diagnosis of poor outcome that needs aggressive treatment) and elaborating individualized treatment plans.

The external clinical assessment and validation of the models, for all four disease areas, will be an on-going process which will benefit from the use of the digital repository and the contribution of clinicians and researchers at various points-of-care. Through the digital repository providers will be able to search for similar clinical cases of children with different diseases and pathological conditions and to compare the predictive model to the patients observed at the clinical setting. This process will increase significantly the number of clinical observation of the repository, thus increasing the accuracy and stability of the model and reducing the tolerance bounds for model prediction.

6. Outline of the clinical assessment and validation for the four disease areas

Models will be assessed with regard both to their clinical accuracy and to their potential contribution to diagnosis support, in particular about:

- Patient-specific simulation
- Patient-specific prediction
- Outcome analysis

Furthermore, in performing the evaluation for each disease area, the specific tables implemented within D2.1 (Initial requirements analysis document including priorities for the implementation) will be taken into account. The validation process will try to check the responsiveness of the models to the specific requirements listed in these tables.

6.1 Cardiomyopathy model

As detailed in WP3 and WP8 the heart model derived from the series of 180 children with CMD will be validated.

The internal clinical assessment of the heart model will test the ability of the derived model to simulate the disease and to predict the evolution of disease and the specific effect of treatment. Specific outcome variables that will be tested are detailed in the table below and will include among others: change over time in functional NYHA class, ability to predict primary endpoints including patient death of need of mechanical support/heart transplant, as well as the evolution over time in a number of specific cardiac features including: left ventricular systolic/diastolic diameters, left ventricular mass, left ventricular volume, left ventricular sphericity index, left ventricular ejection fraction, left ventricular stroke volume, left ventricular mitral valve dimension, mitral valve regurgitant fraction, mitral inflow early and late velocities, mitral tissue Doppler velocities, right ventricular tricuspid plane excursion, right ventricular ejection fraction, right ventricular systolic pressure, intra-ventricular systolic, synchronicity, inter-ventricular systolic synchronicity, Inter-ventricular interaction index, percent regional systolic myocardial deformation.

The external validation is an on-going process which benefits from the use of the digital repository and the contribution of clinicians and researchers at the single points-of-care. At current the digital repository for paediatric cardiology contains medical data acquired in the last 4 years at the OPBG hospital, including clinical data, echocardiographic data and MRI data, comprising more than four million data values. Through the digital repository providers will be able to search for similar clinical cases of CMD children with or without overt cardiac dysfunction and compare the predictive model to the patients observed at the clinical setting. Providers will compare the developed model to the observed patients with regard to: cardiac dimensions, parameters of cardiac systolic and diastolic function, hemodynamic variables as well as to clinical and biochemical characteristics.

Table 1. Outline of Validation for WP-3

Patient-specific features:	Model Testing	Validation Outcome
Patient Simulation:	-Establish patient NYHA functional class -Simulate ventricular volume, ejection fraction and stroke volume. -Reproduce contraction time and ventricular systolic synchronicity -Reproduce systemic pressure curves.	

Patient Disease Progression and Outcome <u>Prediction</u>: - **Effect of Time** - **Effect of therapy:** **a) change in cardiac load** **b) change in heart rate**	Predict changes in: Left ventricular systolic/diastolic diameters Left ventricular mass Left ventricular volume Left ventricular sphericity index Left ventricular mitral valve dimension Mitral valve regurgitant fraction Mitral inflow early and late velocities Mitral tissue Doppler velocities Right ventricular tricuspid plane excursion Right ventricular ejection fraction Right ventricular systolic pressure Intra-ventricular systolic synchronicity Inter-ventricular systolic synchronicity Inter-ventricular interaction index Percent regional systolic myocardial deformation	

6.2 Obesity model

The models produced for obesity patients are complex (including cardiac models, fat distribution models, association rule models, and data intelligence predictive models) and clearly require different approaches to validation. The cardiac models developed within WP9, Task 9.1, will be validated by ground truth (GT) data such as 12-lead ECG measurements and GT anatomy annotations provided by clinical experts, and also with the sensitivity analysis, in order to understand the model's sensitivity to the peculiarities of obese patients. The fat estimation models, Task 9.2, are largely based on segmentation approaches, and are being and will be largely validated by GT annotations provided by clinical experts. The overall data integration and data intelligence tasks (association rule mining, clustering and predictive modelling, Tasks 9.3 and 9.4) will be validated by proper statistical techniques such as bootstrap estimates and cross validation, as already described in the introductory sections of this document. Follow-up data and longitudinal measurements will be involved whenever possible, and, of course, qualitative analysis of the trends and knowledge discovered with a thorough check and analysis by participating physicians will be conducted.

The multi-physics cardiac models will be validated, similar to the cardiomyopathy models described in the previous section, on the GT anatomy segmentation data, reviewed by the clinical partners, and with the ECG data (for the electrophysiology components). In addition, the models will be validated with sensitivity analysis, taking into consideration potential abnormalities of obese patients. From the meetings and discussions between the technical and the clinical partners within the consortium, a list of specificities were identified as important for generation and validation of disease-specific models that will be personalised to the patients of the present study, along with strategies to include them (please see Deliverable 9.1 for more details):

1) Anatomy:

- Increased atrial and ventricular blood volume and ventricular muscle mass: should be obtained and validated through the anatomical segmentation of patient images;
- Sub-epicardial adipose tissue: if images allow segmenting it, it will then have to be introduced as a different tissue type in the model, with different electrical and mechanical properties.

2) Electrophysiology

- ECG displacement by elevated diaphragm. Axis shift because of hypertrophy and elevated diagram. Increase of fat will change the distribution of surface ECG: should be obtained by segmenting the position of the heart within the torso;
- Increased distance between heart and electrodes: should be given by the torso segmentation;
- Increased heart rate, PR, QRS and QT intervals: the electrophysiological parameters controlling these indices can be adjusted if needed.

3) Mechanics

- Increased cardiac output: the increased blood volume should give result an increased output of the ejection fraction. If required, volume overload can be specifically modelled by changing atrial model parameters;
- LV diastolic dysfunction, stiffening: the stiffness parameter will be estimated from the dynamic images.

4) Haemodynamics

- Hypertension: this can be included through the boundary conditions of the model;
- Increased peripheral vascular resistance or decreased (both are seen): this can also be adjusted through the boundary conditions of the model.

Some areas will require more model adaptation and additional validation, in particular the pathological changes at cardiac cell level (e.g. changes in action potential formation and propagation due to fat), the impact of increased stiffness on end-diastolic pressure, and the effect of myocardial fat on the cardiac function. Further areas for additional attention in clinical validation include modelling of the systemic and pulmonic circulation as the left and right parts of the heart are haemodynamically separated in the current model.

The fat estimation models, Task 9.2, are largely based on machine learning-based image segmentation approaches, and are being and will be largely validated by GT annotations provided by clinical experts. In particular, the following measures will be used to validate the techniques; Relative Volume Difference (RVD), Volume Overlapping Error (VOE), Average Surface Distance (ASD), Dice Coefficient (DC), Jaccard Index (JI), Root Mean Squared Error (RMS), and Hausdorff Distance (HD). Table x introduces a sample validation with a cohort of 53 obese young adults from a retrospective study with the above mentioned measures with segmenting the liver in fat-water separated MR images, including minimum, maximum, average and standard deviation for each measurement.

Table 2. Validation measures for automatic liver fat extraction

	Min	Max	Avg	SD
VOE	17.92	52.91	**26.07**	7.07
DC	0.64	0.90	**0.85**	0.05
JI	0.47	0.82	**0.74**	0.07
RVD	-49.13	51.68	**4.24**	17.00
ASD	2.66	15.00	**4.98**	2.17
RMS	4.78	24.55	**8.30**	3.40
HD	18.55	107.39	**41.35**	15.44

The overall data integration and data intelligence tasks (association rule mining, clustering and predictive modelling, Tasks 9.3 and 9.4) will be validated by proper statistical techniques such as bootstrap estimates and cross validation, as already described in the introductory sections of this document. The principal aim of the study, that will be targeted and validated in Tasks 9.3 and 9.4 is to establish patterns of risk factors that best predict the hard intermediate outcomes associated with frank pathology in later life. In order to achieve this, hard intermediate outcome measures will be elaborated and properly validated with the acquired data. This includes already well-established measures as evidence of early disease; measures of insulin resistance, raised blood pressure or increased left ventricular muscle mass. In addition, other factors that normally cannot be measured non-invasively, even with the advanced cardiovascular imaging techniques, will be investigated and validated. In particular, paediatric heart models will be adapted to obtain a greater understanding of the electromechanical function of the heart in our subjects. This should yield estimates of parameters such as ventricular wall strain or wall compliance, for example that would be expected to be abnormal well before the development of hard intermediate outcomes such as ventricular hypertrophy. These parameters will be incorporated in the global predictive model for development of obesity and CVD risk, and validated, in particular paying especial attention to their correlation with already established risk factors.

In addition, the generated predictive models and associations found will be validated against existing studies. In particular, existing cross-sectional studies have been able to link childhood obesity with established surrogate markers for CVD, such as atherosclerosis and cardiac hypertrophy. For example, the Strong Heart Study, which analysed data from over 450 adolescents, demonstrated that patients with obesity and/or metabolic syndrome had a significantly higher prevalence of left ventricular (LV) hypertrophy and left atrial (LA) dilation, together with impaired systolic and diastolic ventricular function.

To ensure quality and utility of the models and workflows developed, a number of procedures will be used. Their performance should be accordingly validated (please see Deliverable 4.1 “Data collection protocol and ethical clearance” for more details):

- Data pre-processing; data validation, discretization, and outlier removal will be performed using the system “Data Curator & Validator” (DCV) developed in “An integrated platform for European paediatrics based on a Grid-enabled network of leading clinical centres” (Health-e-Child) project, and currently being enhanced by T15.1 in MD-Paedigree. Performance of this tool will be validated as part of the MD-Paedigree Infostructure environment, under WP17. An indirect validation of performance of this tool is the improved predictive performance of the models developed in the study after the data curation and validation is accomplished;

- Mapping: descriptors will be cross-mapped to standard data dictionaries (e.g. epSOS value sets, ICD-10 diagnosis, LOINC labs...);
- Data enrichment: normal values will be added from legacy sources (e.g. clinical guidelines);
- Data description: standard descriptive statistical tests will be applied for categorical and for real-valued variables; distance measures will be used to facilitate case-based retrieval; clustering of cases will be carried out using a number of advanced statistical & visualization approaches; visual data will be aggregated with textual and other structured data to facilitate differentiation of normal from abnormal visual data; simulations will be performed based on Bayesian network-based knowledge discovery techniques, using the AITION (Scalable Platform for Interactive Data Mining) system, developed in the Health-e-Child project and further extended by T16.3 (see also the later section on AITION and WP16.)

Appropriate methods will also be used to control the statistical power of the study and validate the presence of statistically significant differences between obese and non-obese participants, and also to differentiate between the different degrees of obesity, for the cardiovascular, endocrine, metabolic and inflammatory measures. These methods will include but not be limited to univariate and multivariate linear regression and group mean comparisons between obese and non-obese participants using t-tests or appropriate non-parametric equivalents. Where possible, non-normally distributed variables will be transformed to normality prior to parametric testing. All the findings drawn in the study will be validated with statistical significance tests, similar to the other studies in the project. These include parametric tests, whenever appropriate, such as the Student's t-test, and, in other cases, non-parametric tests such as McNemar's test for difference of two proportions.

6.3 Juvenile Idiopathic Arthritis (JIA)

Juvenile idiopathic arthritis is a group of disorders characterized by chronic joint inflammation. Uncontrolled inflammation can lead to joint damage, manifesting itself as deformities and irresolvable functional disability. Therefore, the goal of treatment is to induce early disease remission, preventing the onset of joint damage.

The biomechanical ankle model will be used to simulate lower-limb movement and mechanics during a full gait cycle and to predict the onset and progression of joint damage in the lower–limb, specifically at the ankle. At the first stage of model assessment and validation, each individual model will be tested to evaluate if it is able to simulate the lower limb movement and mechanics. Specific attention will be paid to the following parameters: joint angles (using comparison to clinical gait analysis data), inter-articular forces (using comparisons to *in vivo* data, acknowledging the associated limitations) and muscle forces (using comparison with electromyography data collected in parallel).

Next, the models' ability to predict patient outcome will be tested with internal validation. The outcome measures that will be used are joint damage, measured using magnetic resonance imaging and the juvenile arthritis damage index (JADI); disease remission according to internationally validated criteria; and functional

ability, measured using the childhood health-assessment questionnaire (CHAQ). A large number of model outputs will be analysed with the aim of finding a relationship to disease progression. Some parameters that are hypothesised to have a close relationship with disease progression are: peak tibio-talar cartilage stress, inter-articular forces, specific muscle's forces, ground reaction force, foot contact time, gait symmetry and inverse dynamics moments. These simulated parameters, as well as other parameters that are found to have a relationship to disease progression or parameters that can be used as simple forecasters of more complex parameters, will be used as predictive parameters. Internal validation will take place using the patients provided by MD-Paedigree and will test whether the model is able to predict disease outcome (in terms of the abovementioned outcome measurements) for patients with different categories of JIA and differing severity of the disease.

The external validation will take place once the repository is populated with external data. The model will then be validated in a similar way as during internal validation, using these external data.

6.4 Neurological and Neuromuscular Diseases (NND)

As detailed in WP6 and 11, lower extremity musculoskeletal models will be developed for 30 children with CP, 20 children with DMD, and 20 children with CMT diseases. Furthermore, statistical, probabilistic models will be developed using ~300 retrospective pre-post treatment cases for CP, which will allow individualized treatment outcome prediction. Furthermore, for DMD and CMT, longitudinal data collection and storage in the repository will allow identification of biomarkers to predict the progression of the disease over time. Below, first the internal validation of the developed biophysical models will be outlined, followed by the external validation of treatment outcome predictions.

The internal validation of the biophysical models that allow patient-specific simulation is twofold. First, two levels of model complexity are implemented and compared. As a first level of detail, models will be personalized based on relatively easily collectable clinical data, such as standardized physical exam, anthropometry measurements and functional joint calibration. Furthermore, models will be personalized on a more in depth detailed level using outcomes of MR imaging, which can be considered a 'gold standard' for several parameters (e.g. bone shapes, muscle volumes). The two types of models (clinical data-based and imaging data-based) will be compared to identify the level of detail that is clinically relevant and needed for accurate individualized mode predictions. This comparison between clinical data and image data-based models is accompanied by extensive sensitivity studies providing insight in the clinical relevance of model parameter adjustments.

Second, as a more direct validation, several key features of the two types of patient-specific musculoskeletal models (clinical data-based and imaging data-based) and their accompanying gait simulations are compared

to experimental data not used to develop the models. The table below gives an overview of main patient-specific model features and how they will be validated.

Table 3. Outline of Validation for WP-6/11 (w.r.t. Patient-specific simulation)

Patient-specific features	Model Testing	Validation Outcome
Maximum joint torques	Simulate joint torque at maximum (100%) activation	Simulated maximum joint torque matches experimentally measured maximum joint torque
Muscle activity during gait	Simulate the measured gait data and optimize the muscle activity responsible for the observed movements, using a muscle force-sharing cost function	Simulated muscle activity matches experimentally measured EMG
Total muscle work / energy	Calculate the total mechanical work and energetic equivalent summed over all muscles	Simulated total work / muscle energy should correlate with experimentally measured energy expenditure

External validation of probabilistic model predictions is an ongoing process which benefits from the use of the digital repository and the contribution of clinicians and researchers at the clinical centres. Currently we are in the process of selecting retrospective data of as many as possible pre- and post-treatment cases, aiming at 300 pre and post treatment combinations for CP at month 36, and ongoing increasing after that. This data will be used for predictive statistical model development, which will be validated by using only part of the data to develop the statistical models, and testing it on the left-out cases. For CP, the predicted treatment outcome of these left-out cases will be compared to actual treatment outcomes, while for CMT/DMD the predicted progression of the disease will be compared to the actual progression as observed in the left-out cases. The statistical model predictions are expected to improve with increasing number of cases in the repository, increasing consistency of the (prospective) data, and increasing number of relevant biomarkers calculated using biophysical modelling.

7. Usability evaluation through the Infostructure

Ultimately, the final usability evaluation of the clinical models will be performed through the MD-Paedigree Infostructure. As mentioned in section 3.2, the e-infrastructure applications supporting the clinical models released by the Infostructure will follow a Scrum agile process, with a view on user validation and interfaces quality improvement. Under an agile approach, requirements, programming, and testing are often done concurrently. Agile development recognizes that testing is not a separate phase, but an integral part of software development, along with coding. Testing and coding are done incrementally and iteratively, building up each feature until it provides enough value to release to production. Within the agile process, control procedures for acceptance testing should be defined before usability testing and user acceptance. Test cases are built around specifications and requirements, i.e., what the application is supposed to do. Visual testing could also be possibly utilised for user acceptance and usability testing, making it easy for users to provide detailed bug reports and feedback (e.g. visual testing can record user actions on screen, as well as their voice and image, to provide a complete picture at the time of software failure for the developer). Testing is used in association with verification (have we built the Infostructure correctly?) and validation (have we built the right Infostructure? i.e. do the requirements satisfy the users?).

Testing and validation of the MD-Paedigree Infostructure will be carried under the tasks of WP17. The objective of WP17 is to test and validate the results of the activities carried out under WP 14, 15 and 16. The deliverables that will be produced by WP17 include test reports on MD-Paedigree's Alfa [month 24] & Beta Prototype [month 36], and the final MD-Paedigree platform [month 48], as well as, test reports of the Case-Based retrieval prototype [month 36], and the beta [month 36] and final release prototype [month 48] of the AITION platform, described in the following section.

7.1 AITION Big Data Healthcare Knowledge Discovery platform (WP16)

To improve the validation process, outcome analysis and personalized prediction MD-Paedigree integrates AITION which is part of WP16. AITION is an information processing, knowledge discovery and simulation platform developed by the MaDgIK Lab[1] (ATHENA RC and University of Athens) that supports the fundamental shift in the way biomedical data are collected and processed, as well as how biomedical research is performed in the area of Big Data Healthcare. Focusing on bottom-up (evidence-oriented) analysis, AITION aims to identify latent factors (disease signatures) and related biomarkers that can explain and predict variabilities in drug therapies, disease evolution and outcome, revealing similarities among patients. At the same time AITION provides reusable patient/disease specific statistical simulation models analysing correlations on high dimensional, heterogeneous, multi-modal biomedical data promoting decision support, outcome analysis and model-guided personalized medicine.

In more detail, AITION supports an end-to-end, pre-processing/validation, information retrieval, knowledge, pattern and similarity discovery, and eventually decision support, simulation and reasoning workflow. Based on this, the researcher has the ability to define multiple experiments - consisting of several data analysis steps - that can be incrementally updated as more data enter the system. Such experiments, maintain provenance

[1] www.madgik.di.uoa.gr

and related analysis information leading to well documented, reproducible results. Proposed workflows consist of the following well-defined steps/tasks:

- **Data Preprocessing and Validation:**

Pre-processing is the first step in the workflow, ensuring that the raw data is validated, curated and appropriately transformed (e.g. discretized) for use by the data analysis system. AITION provides advanced techniques for data curation, validation and pre-processing, via the DCV tool developed under T15.1. It is able to: validate data attributes; check for outliers, inconsistencies and mistakes in the data; automatically resolve many of the frequently missing/null values; compute medical scores, aggregates, and other derived variables; and perform attribute discretization when required. AITION's curation mechanisms are based upon rules that can be modified/ added by the user.

- **Disease Signature Identification, Patient Grouping and Similarity Detection**

AITION provides advanced multi-modal, mixed-membership clustering and similarity analysis techniques, extending well established unsupervised techniques, as well as classical Bayesian parametric and non-parametric latent variable models, mixtures and ad-mixtures targeting disease signature identification.

Disease signatures: refer to latent factors (patterns) that characterize the disease, as well as related differentiations. In AITION, such patterns are described as distributions on the most relevant variables per disease (e.g., biomarkers). Such variables commonly only refer to expressions of the disease in terms of subjective (symptoms, clinical evaluation) and objective (labs, genes, proteins) examination – and sometimes may include existing assessment characterization – and do not usually include therapy and intervention. Nevertheless, intervention should be taken into consideration since it likely affects the expression of the disease.

Disease signature allocation on patients: we address this problem following a mixed membership approach. This means that a patient is not characterized solely according (or belongs) to a single disease signature (hard clustering), but rather has a specific distribution (proportion) on more than one disease signatures. Thus, we have multiple variables per signature and multiple signatures per disease.

Patient Grouping and Similarity: based on the above-mentioned patient-to-disease signature allocation, we are able to identify homogeneous groups among patients and model case-based patient similarity (patients like me or mine) based on several metrics (such as cosine similarity, symmetric Kullback–Leibler divergence and others).

- **Statistical Simulation Modelling**

Having identified homogeneous groups of patients according to disease signatures, we go one step further towards personalized medicine, building highly accurate and reusable predictive – patient or disease specific – statistical simulation models combining a:

- bottom-up data-driven process, i.e. analysing heterogeneous, vertically-integrated demographic, phenotypic, clinical, molecular and genomic bio-medical data, images and streams; and a

- Top-down model/concept-driven process, i.e. incorporating external knowledge coming either from domain experts, literature, or model-driven processes and relational/semantic models.

Thus, by combining disease signatures' analysis with other learning procedures that reveal important correlations between variables in all dimensions and scales, we can build statistical simulation models that include therapy planning (possible treatment plans, medications, etc.), as well as outcome. Such models could be validated and improved later through retrospective analysis.

AITION provides feature selection, variable correlation analysis, statistical simulation modelling and reasoning based on Graphical Probabilistic Models (GPMs). GPMs are a popular and well-studied framework for compact representation of a joint probability distribution over a large number of interdependent variables, as well as for efficient reasoning about such a distribution. AITION implements state-of-the-art algorithms and techniques for Bayesian Network (BN) Structure and Parameter Learning, Markov Blanket induction and feature selection, and real-time inference. Furthermore, ontologies and a-priori knowledge can be incorporated with the BN, defining topological constraints, in order to automate causal discovery and feature selection and provide semantic modelling under uncertainty. This way, AITION presents a rich 'natural' framework for imposing structure and prior knowledge, providing the domain expert with the ability to seed the learning algorithm with knowledge about the problem at hand.

- **Personalized Model-Guided Medicine and outcome analysis**

The ultimate goal of this task is to support personalized disease prediction and prognosis, as well as adapted and individualized therapy plans. Utilizing statistical simulation models, a clinician or researcher can analyse multiple "what if" scenarios supporting decision making. In addition, such models can support or validate already defined clinical workflows and be used for outcome analysis.

7.2 Case-base retrieval

The case-base retrieval engine is developed in close interaction with the end-users. The developments are driven by three user channels: user requirements (WP13), relevance judgement by experts (WP15), and ultimately tests (WP17). WP13 first specifications have already been delivered (D13.1) together with the first benchmark-based evaluation of the case base retrieval engine (D15.1).

The core search engine is based on a vector-space model, powered with a normalized set of onto-terminological descriptors (Medical Subject Headings). The preliminary evaluations performed in D15.1 suggest that retrieval effectiveness has already reached a precision at high ranks (similarity of the top returned case) sufficient to support the exploration of similar cases: today more than 7 out of 10 requests for similar cases are able to retrieval a relevant patient file. The rest of the foreseen evaluation tasks are progressing as planned, in particular the querying instruments to assess the population corresponding to a particular profile (clinical case and/or trial protocols) in the data repository.

8. Conclusions

This document represents the first step of the process of clinical validation of the models and Infostructure which will be implemented in MD-Paedigree. The assessment process will be composed of three subsequent documents on clinical assessment and validation results for all four disease areas (at M24-36-48). If successful, the process of validation will eventually lead (at M48) to D12.3, “Improved clinical workflows and outcome analysis: Final proposal of innovative clinical workflows based on outcome analysis of all patient cases”, which will represent the core of the project in terms of clinical outcomes and actual innovation in the everyday clinical practice.