EU-STANDS4PM
standards for in silico models for personalised medicine

A European standardization framework for data integration and data-driven in silico models for personalized medicine – EU-STANDS4PM

EU-wide mapping report on good practice examples for integrating phenotype and large scale data

Deliverable 1.2

Table of contents

Introduction

One of the major goals of EU-STANDS4PM is to assess and evaluate national standardization strategies for interoperable health data integration as well as data-driven in silico modelling approaches for personalized medicine with the aim to bundle European standardization efforts. The project aims to produce an in depth EU-wide mapping of relevant European initiatives with regard to data sources (work package 1) as well as in silico models (work package 2). This process is the foundation to assemble specific recommendation and guidelines (and EU standard documents) for data harmonization/integration strategies as well as data-driven in silico approaches to interpret human disease/health data. One of the challenges is to establish which standards are already in use and where there are gaps, and to analyse how these affect existing projects. To this aim, we assembled a group of excellent projects utilising in silico models for the purposes of enabling personalised medicine, to examine the current state of the art together with gaps and wishes for the future.

EU-STANDS4PM is an open network and seeks input from all relevant stakeholders that have an interest in advancing predictive in silico modelling approaches in personalized medicine – through broadly applicable standards and harmonized procedures for integration and interpretation of human disease/health data.

Good practice examples for integrating patient derived data

The current report was initiated by an international stakeholder consultation event that took place on 4 February 2020 at the Heidelberg Institute for Theoretical Studies, Germany. The event was organized as a joint effort between work package 1 “Data sources and standards for predictions in personalized medicine” and work package 2 “Integrative data analysis and in silico models in personalized medicine”.

The focus of the workshop was on national and EU-case studies (good practice examples) for integrating patient derived data, such as phenotype and large scale data, for in silico modelling in personalized medicine. In the context of personalized medicine, the use of patient-derived data for disease-related modelling was discussed, as well as the potential use of artificial intelligence to utilize the patient-derived data and support the modelling process. Thus, major aspects of the workshop were to:

- **Reflect** on specific use cases of using patient-derived data for modelling (legal aspects are discussed in two accompanying reports available on www.eu-stands4pm).
- **Discuss** the potential use of artificial intelligence methods to utilize patient-derived data for modelling (e.g. to integrate, summarize and/or visualize large data sets).
- **Map and compare** state-of-the-art modelling approaches (manual mechanistic/systems medicine approaches) based on patient-derived data with future possibilities for more automated AI-based processes to analyse large scale and distributed datasets from patients.
- **Analyse** needs-and-gaps to outline and develop a standardization roadmap for the use of AI technologies and manual modelling approaches in the context of in silico models for personalized medicine

In the following section the outcomes of the workshop are summarized as a collection of key topics that were further elaborated and discussed in greater detail in a White Paper (Deliverable 2.3, submitted), an ISO Technical Specification (ISO/AWI TS 9491), as well as a review article (currently under peer review, Journal of Personalized Medicine):

Clinical and modelling language are different and key is linking them interoperably through automated mapping ontologies

- What does model validation mean to different people?

Importance of dynamic prediction and of model explainability

- Which principles should drive model validation?

- How do we ensure data harvestability at source and avoid problems of data hostage-taking by software suppliers?
- Which principles should regulators use to establish whether researchers understand how their models work?

Model validation, credibility assessment, good simulation practices

- Model validation should be as generic as possible (agnostic to model specificity)
- How do we bring models that work, back to the patient?
- Outcome measures that measure patient well-being – How do we define a gold standard (e.g. mortality reduction, Patient well-being)?

Input data, model validation, and EU leadership

- How to use Artificial intelligence (AI) vs. manual approaches for modelling in personalized medicine (potential use of AI methods vs. manual approaches to utilize patient-derived data for modelling (e.g. to integrate, summarize and/or visualize large data sets) and the specific standardization needs to do that)?
- How to implement a standardization roadmap for the use of AI technology and manual in silico modelling approaches for personalized medicine?

Standardization efforts for EU-STANDS4PM and mechanism to use

- Development of ISO standards for computational models.
- Data processing and integration includes quality management of data and models.
- Tracking data provenance.
- Minimum reporting standards for models
- Draft Technical specification or
- Draft technical report for ISO
- Important for validation

The broad discussion of these topics generated a better understanding of the challenges associated with integrating patient derived data for in silico modelling in personalized medicine. A sense of how enormous the scope is, was confirmed, and all participants were reminded of how many people are working to address such issues, and how important it is to avoid duplication of work without progress.

Motivation to develop standards for data integration

Major challenges in the field of personalised medicine are to harmonize the standardization efforts that refer to different data types, approaches and technologies, as well as to make the standards interoperable so that the data can be compared and integrated into models. Reproducible modelling in personalised medicine requires a basic understanding of the modelled system, as well as of its biological and physiological background. There is a relevant checklist that provides guidelines on the minimum amount of metadata information required in order to understand a model (minimum information requested in the annotation of biochemical models (MIRIAM), (Le Novere, Finney et al. 2005)). This information about data and models can be transferred by using metadata in the form of semantic annotations. These annotations can improve the shareability, and interoperability of the data or model (Neal, Konig et al. 2019). To render data and models FAIR, it is important that all their elements (entities) in their context are understood in exactly the same way, independently from the individual or tool that process or analyses them. For this purpose, it is necessary to consistently use the defined terminologies, such as controlled vocabularies and domain ontologies that can be defined and applied independently of the data/model format.

For many different data types used in personalised medicine domain-specific annotation standards and terminologies are available. For example, UniProt[1] or the Protein Ontology[2], can be used to uniquely identify proteins in a particular biological context which can then be linked to specific entities in the computational model. Similarly, the Gene Ontology[3] could be used to identify specific genes or cellular components whereas the Foundational Model of Anatomy (FMA) (Rosse and Mejino 2003) can be used to localize an entity in the computational model to specific spatial location or anatomical structure. If not found completely or partially unstructured, which is often the case, health-related data is most commonly structured and codified by specific formatting standards for medical data. These can be the interoperability standard HL7 Fast Healthcare Interoperability Resources (FHIR) (Bender and Sartipi 2013), or the standard for electronic health records (openEHR (Kalra, Beale et al. 2005)). Semantical content is usually annotated with domain-specific clinical terminologies, e.g., International Classification of Diseases (ICD)[4], Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT, see table 1, annex), or Logical Observation Identifiers Names and Codes (LOINC, see table 1, annex). Thus, the wheel does not have to be reinvented for the semantic data annotation in personalised medicine, but existing annotation standards have to be consistently applied.

Methods of integrating data for personalized medicine

Data can be integrated in different ways, with different levels of personal identifiability.

Individual level integration

Research and clinical data can be linked at the personal ID level so that research-generated and clinical care data become a combined pool of knowledge about the given individual. As the different data sources contribute different variables, liked to a given individual, this type of research-clinical data integration can be done without mapping data to uniform ontologies. Missingness and contradictions can be handled through analysis of the data source and data creation route, and extremely valuable knowledge is generated through these projects.

Integration of variables

Data can be combined at the variable level, e.g. diagnoses may be integrated from national records using ICD-codes, projects using project-based definitions of disease, and questionnaires using self-identification by patients. Blood sugar values may be combined from national patient journal lab records, project lab data with different equipment and normal values, and patient home self-measurements. This obviously requires large mapping efforts to standardised mapping concepts and produces unreliable data integration and results.

Solutions used in the integration of large-scale, transnational project data from multiple clinical centres could be used to address the above, e.g. calibration of lab values to handle inter-lab variations and techniques to impute missing values. Especially useful are experiences of integration of research project data from projects arising through of combining different, existing projects, where project SOPs and definitions are dissimilar.

Integration of unstructured data

Unstructured, non-standardised data elements such as journal free text can be text mined through Natural Language Processing and the creation of specific ontologies and dictionaries to extract relevant information. Interestingly the rawness of these data makes them interoperable, in the sense that similar dictionaries can be created and similar rules applied to text from different sources and the results are

[1] UniProt: https://www.uniprot.org/;
[2] Protein Ontology: https://www.ebi.ac.uk/ols/ontologies/pr
[3] Gene Ontology: http://geneontology.org/
[4] ICD: https://www.who.int/classifications/icd/en/

structured in the same way. Images are another example of unstructured data which can in theory, at least, be processed and then integrated.

However research data and clinical data will still have been created under different circumstances and in different patient populations, and the metadata surrounding the data are often inadequate, complication data integration.

Federation of data or validation of findings

Clinical and research data can contribute to joint results by training an algorithm sequentially on the data sets without combining them. Validating results derived from clinical data, using research data (or the other way around) is also a way of avoiding having to combine data sets governed by different ethical, legal and security constraints. However, the data sets still have to be standardised and interoperable to return useful results.

In essence, data generated both by research projects and clinical care should be designed for interoperability

Common standards relevant for personalised medicine

Table 1: Common standards relevant for personalised medicine and *in silico* approaches.
Examples of common standards that have been developed by specific user communities and different stakeholders. Their use has been enhanced as they have been coupled to tools which have spread in the respective field of research. In addition, a current overview about data formats and standards for in-silico systems biology and quantitative modelling can be found in (Golebiewski 2019) and as a comprehensive reference in the annex of ISO 20691 (in preparation).

DNA, RNA, protein sequence formats	
FASTA	Widely used for representing nucleotide sequences or amino acid, developed for use in the FASTA program (Lipman and Pearson 1985, Pearson and Lipman 1988). The FASTA format is simple and lacks facility for extensive annotation.
Sequence Alignment/Map (SAM) and Binary Analysis Map (BAM) format	Capture of sequences that have been aligned to a reference genome. SAM is a tab delimited text format consisting of a header section, which is optional, and an alignment section. BAM is in a binary more condensed version while SAM has the same information in a series of tab delimited ASCII columns (Li, Handsaker et al. 2009). BAM files are compressed files.
CRAM	A compressed columnar file format also used for storing biological sequences mapped to a reference sequence, it has been developed to improve compression and hence save on storage costs (Hsi-Yang Fritz, Leinonen et al. 2011).
ISO/IEC 23092 (MPEG-G): Information technology — Genomic information representation	The ISO/IEC 23092 (MPEG-G) series of standards is a coordinated international effort to specify a compressed data format that enables large scale genomic data processing, transport and sharing. Interoperability and integration with existing genomic information processing pipelines is enabled by supporting conversion from/to the FASTQ/SAM/BAM file formats. It consists of currently (as of October 2020) six parts: Part 1: Transport and storage of genomic information Part 2: Coding of genomic information Part 3: Metadata and application programming interfaces (APIs) Part 4: Reference software Part 5: Conformance Part 6: Coding of genomic annotations

General feature format (GFF)	Stores DNA, RNA or protein genetic sequence data (Akanksha Limaye 2019). It stores the whole sequence for the relevant feature.
Variant call format (VCF)	A text format file storing the same data but only contains the sites which differ from a given reference and hence is more space efficient than GFF (GitHub_Community 2020). Originally designed to be used for SNPs and INDELs but can also be used for structural variation. A Variant represents a change in DNA sequence relative to some reference. For example, a variant could represent a Single Nucleotide Polymorphism (SNP) or an insertion. Variants belong to a VariantSet. This is equivalent to a row in VCF.
Binary variant call format (BCF)	A binary version of VCF and therefore is more space efficient, the relationship between BCF and VCF being similar to that between BAM and SAM.
Synthetic Biology Open Language (SBOL)	An RDF/XML format for representing, among other things, sequences for genetic circuit designs. It has a rich ability to express both sequence feature annotations and part/sub-part relationships. It is also designed to represent incomplete/partial sequences and relative ordering of parts in a genetic design.
Mass Spectrometry	
mzML	Stores the spectra and chromatograms from mass spectrometry in and eXtensible Markup Language (XML) format. Now a well-tested open-source format for mass spectrometer output files that is widely used (Martens, Chambers et al. 2011).
mzTab	A more easily accessible format which could be used with R or Microsoft Excel tools in the field of proteomics and metabolomics. mzTab files can contain protein, peptide and small molecule identifications. In addition experimental meta-data and basic quantitative information (Griss, Jones et al. 2014).
Medical imaging, Digital Imaging and Communications in Medicine	
Digital Imaging and Communications in Medicine (DICOM)	Dominating standard used in medical radiology for handling, storage, printing and exchanges of images and related information. Specifies the file format and communication protocol for handling these files. Captures pixel data making up the image and how the image was generated (e.g., used machine and protocol, information regarding what patient the image is capturing. Living standard regularly maintained and modified (DICOM_Secretariat 2020), also adopted as ISO 12052 "Health informatics - Digital imaging and communication in medicine (DICOM) including workflow and data management".
The European Data Format (EDF)	A standard to archive, share and analyse data from medical time series (Kemp, Varri et al. 1992).
Semantic integrations	

BRIDG (Biomedical Research Integrated Domain Group Model)	An information model being used to support development of data interchange standards and technology solutions to enable semantic (meaning-based) interoperability within the biomedical/clinical research arena and between research and the healthcare arena. BRIDG is a collaborative effort engaging stakeholders from the Clinical Data Interchange Standards Consortium (CDISC), the HL7 BRIDG Work Group, the International Organization for Standardization (ISO), the US National Cancer Institute (NCI), and the US Food and Drug Administration (FDA). The goal of the BRIDG Model is to produce a shared view of the dynamic and static semantics for the domain of basic, pre-clinical, clinical, and translational research and its associated regulatory artifacts. The BRIDG Model is a hybrid of conceptual and logical models represented as UML Class Diagrams. It was built by harmonizing other project and domain models and each concept in the BRIDG model carries its provenance in the form of mapping tags indicating what data elements from other models map to that concept.
HL7 FHIR (Fast Healthcare Interoperability Resources)	A standard for exchanging healthcare information electronically.
Human Phenome Ontology (HPO)	Developed by the Monarch Initiative a consortium, carrying out semantic integration of genes, variants, genotypes, phenotypes and diseases in a variety of species allowing powerful searches based on ontology. HPO is a standardized vocabulary of phenotypic abnormalities associated with disease. Standard terminology for clinical “deep phenotyping” in humans, providing detailed descriptions of clinical abnormalities and computable disease definitions (Shefchek, Harris et al. 2020). The primary labels use medical terminology used by clinicians and researchers. These are complemented with laypersons synonyms. HPO is one of the projects in the Global Alliance for Genomics and Health (GA4GH) seeking to enable responsible genomic data sharing within a human rights framework (GA4GH_Community 2020).
SNOMED CT	The Systematized Nomenclature of Medicine (SNOMED) is a family of medical terminology systems. Originally conceived as a nomenclature, the latest version SNOMED CT can best be characterised as an ontology-based terminology standard. The goal of all SNOMED versions is to provide a language that represents clinical content as clearly and precisely as possible, regardless of its original language. This should enable search queries to be answered with high recall and high precision (see also www.snomed.org).
LOINC	The Logical Observation Identifiers Names and Codes (LOINC) is a database of common names and identifiers used to identify laboratory and clinical examination and test results. The aim is to facilitate the electronic exchange of data when transmitting medical examination results and findings data. LOINC is recommended (also by HL7 and DICOM) for the exchange of structured documents (CDA) and messages (see also https://loinc.org/).
Serial ISO/IEEE 11073	Personal Health Data (PHD) Standards, a group of standards addressing the interoperability of personal health devices (PHDs) such as weighing scales, blood pressure monitors, blood glucose monitors, etc. (see also: http://11073.org/).

Models and modelling tools	
CellML	A standard based on XML markup language (Lloyd, Halstead et al. 2004) used for storing and exchanging computer-based mathematical models allowing sharing of models even when different modelling tools are used (Schreiber, Bader et al. 2016). CellML is a description language to define models of

	cellular and subcellular processes and supports component-based modelling, allowing models to import other models, or subparts of models, therefore strongly encouraging their reuse and facilitating a modularized modelling approach. A CellML model typically consists of components, which may contain variables and mathematics that describe the behaviour of that component. The mathematical model is considered to be the primary data and biological context is provided by annotating the variables and equations with metadata using the Resource Description Format (RDF).
The Systems Biology Markup Language (SBML)	A standard model interchange languages that permits exchange of models between different software tools (Hucka, Bergmann et al. 2018). SBML is a machine-readable, XML (Extensible Markup Language) based model description and exchange format for computational models of biological processes. Its strength is in representing phenomena at the scale of biochemical processes, but it is not limited to that. The evolution of SBML proceeds in stages (levels). Since SBML Level 3 the format is modular, with the core usable in its own right and packages being additional "layers" that add features to the core. SBML core is suited to representing such things as classical metabolic models and cell signaling models. SBML packages that extend the core and are optional in their use, add additional model features, such as visualizations, distributions, constraint-based models (flux balance constraints), hierarchical model composition, special processes or grouping of elements.
The Synthetic Biology Open Language (SBOL)	A standard to support specifications and exchange of biological design information (Madsen, Moreno et al. 2019). SBOL Data provides both an electronic format for representing this information, while SBOL Visual provides schematic glyphs to graphically depict genetic designs.
Simulation Experiment Description Markup Language (SED-ML)	A machine-readable, XML (Extensible Markup Language) based format for encoding the description of a computational simulation. Developed to capture the Minimum Information about a simulation experiment (MIASE), the minimal set of information needed to allow reproduction of simulation experiments (Waltemath, Adams et al. 2011, Schreiber, Sommer et al. 2019). Typically used with an XML-based model description format (e.g. CellML or SBML), SED-ML allows for the description of applying a numerical algorithm to a mathematical model in order to perform a given task. Tasks may be nested to allow the composition of relatively simple tasks into increasingly complex simulations. Mechanisms exist in SED-ML to apply pre-processing steps to a model prior to executing a simulation task and also to apply post-processing to the raw simulation results (Nickerson et al. 2016).
Open Modelling EXchange format (OMEX)	OMEX supports the exchange of all the information necessary for a modelling and simulation experiment in the life sciences. An OMEX file is a ZIP container that includes a manifest file, an optional metadata file, and the files describing the model. The manifest is an XML (Extensible Markup Language) file listing all files included in the archive and their type. The metadata file provides additional information about the archive and its content. Although any format can be used, an XML serialization of the Resource Description Framework (RDF) is recommended (Bergmann, Adams et al. 2014).
NeuroML	XML-based standardized model description language to describe mathematical models of neurons and complex neuronal networks (Goddard, Hucka et al. 2001). The focus of NeuroML is on models which are based on the biophysical and anatomical properties of real neurons.
PBPK/PD	Physiologically based Pharmacokinetic/Pharmacodynamic models allow a mechanistic representation of drugs in biological systems (Kuepfer, Niederalt et al. 2016).
Pharmacometrics Markup Language (PharmML)	A machine-readable, XML (Extensible Markup Language) based model description and exchange format used for encoding computational models, associated tasks and their annotation as used in pharmacometrics. It provides

	the means to encode pharmacokinetic and pharmacodynamic (PK/PD) models, as well as clinical trial designs and modelling steps (Nickerson, Atalag et al. 2016).
Human Physiome Field Markup Language (FieldML)	A machine-readable, XML (Extensible Markup Language) based model description and exchange format for representing hierarchical models using generalized mathematical fields. FieldML can be used to represent the dynamic 3D geometry and solution fields from computational models of cells, tissues and organs (Nickerson, Atalag et al. 2016).
Biological Pathways Exchange (BioPAX)	A machine-readable standard format that aims to enable integration, exchange, visualization and analysis of biological pathway data.
Numerical Markup Language (NuML)	A machine-readable, XML (Extensible Markup Language) based format for describing and exchanging multidimensional arrays of numbers to be used with model and simulation descriptions.
Analysis pipelines	
ISO 25720	Genomic Sequence Variation Markup Language (GSVML). The standard is applicable to the data exchange format that is designed to facilitate the exchange of the genomic sequence variation data around the world, without forcing change of any database schema, based on XML.
ISO/TS 20428	Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records. The specification defines the data elements and their necessary metadata to implement a structured clinical genomic sequencing report and their metadata in electronic health records particularly focusing on the genomic data generated by next generation sequencing technology.
ISO/DIS 21393 (in preparation)	Omics Markup Language (OML). OML is a data exchange format designed to facilitate exchanging omics data around the world without forcing changes to existing databases.
ISO/DTR 21394 (in preparation)	Health informatics — Whole Genome Sequence Markup Language (WGML)

Case Study Collection

The following sections contain a collection of case studies of various collaborative research projects that successfully used patient derived data and predictive modelling in a clinical set up.

Project	**BD2Decide: Big Data and Models for Personalized Head and Neck Cancer Decision Support (http://www.bd2decide.eu/)**
Methods	a) RNA: STAR/DeSEQ2 b) DNA methylome: Bismark/RnBeads c) Microbiome: Mothur/Qiime/Humman2/Metaphlan d) Marker identification and reduction to diagnostic sets are done by different ML approaches (random forest , Bayesian inference)" e) Machine learning (unsupervised) f) Omics data analysis g) Omics data analysis
Model function	a) Combined methods making survival prediction based on clinical factors in HNC process
Input data	a) Clinical and pathological b) Clinical and pathological c) Clinical, pathological, genomics (transcriptomics) and radiomics d) Clinical, pathological, genomics (transcriptomics) and radiomics e) Tissue samples f) Imaging data
Output	a) Survival prediction Purpose: To assess the impact, in terms of survival, of each clinical factor involved in the HNC process. b) Patient/cohort classification. Purpose: To identify relations among different variables that apparently are not related. c) Patient/cohort classification Purpose: To identify variables that are correlated to a certain group of population (new patterns). d) Patient/cohort classification Purpose: To uncover significant indicators associated to patient cohorts. e) Transcriptomic profiling Purpose: To deal with high amount of genes and discover relations between genes and patient cohorts. f) Radiomics profiling Purpose: To deal with high amount of radiomic features and discover relations between radiomics and patient cohorts.
Processing steps	a) Univariate (Log-rank test), Multivariate Cox Model, Survival trees b) Logistic regression, Support Vector Machine, Random Forest c) K-means, Birch, Ward, Spectral Cluster d) Principal Component Analysis, Independent Component Analysis, Non-negative Matrix Factorization e) RNAseq (next-generation sequencing) f) Radiomic feature extraction

Use cases	a-f) Patients affected by TNM stage III-IV head and neck cancers
Scale (tissue, organ, cell etc.)	a) All b) All c) All d) All e) Sub-cellular & Tissue scale f) Tissue scale & Organ scale
Challenges/Benefits/Limitations, Input Data standardisation	a) In order to avoid issues on data standardization, due to data comes from different hospitals, well-defined protocols and data cleaning processes have been adopted, inspired also in previous works. b) Laboratory problem were solved centralizing genomic tumour tissue samples analyses in only one clinical centre. c) Technical approaches has been applied to solve the problem of sharing data in various centres and to compare the data with other projects (like RARECAREnet). Dedicated and private services, and integrated approaches has been applied. d) IT restrictions in each hospital to share and collect the data were addressed to allow the usage of secure services in each centre. e) Ethical and data protection regulation has been followed-up to allow the correct use of the data management during the project execution.
Existing standards (formats, guidelines, ontologies)	a) BD2Decide ontology (mapped with external ontologies such as SNOMED-CT or ICD10). Also mapped with Gene ontology. b) Ontology was based on previous project (NEOMARK). c) Within the Decision Support System, as part as the Knowledge Management System, a set of rules has been defined to be used in future projects as guidelines. d) Ethical issues force to create informed consent to obtain the authorization by Ethical committee in each centre, complying each legal standard.
Model validation	no information available
Lessons and Comments	no information available

Project	**Computational Horizons In Cancer (CHIC): Developing Meta- and Hyper-Multiscale Models and Repositories for In Silico Oncology (http://chic-vph.eu/project/)**
Methods	no information available
Model function	-Adapting a Four Dimensional Nephroblastoma Treatment Model to a Clinical Trial Case Based on Multi-Method Sensitivity Analysis (Georgiadi, Dionysiou et al. 2012) -The Technologically Integrated Oncosimulator: Combining Multiscale Cancer Modelling with Information Technology in the In Silico Oncology Context (Stamatakos, Dionysiou et al. 2014).
Input data	Tomographic imaging data, clinical data, molecular data, pathology data
Output	Response to neoadjuvant treatment
Processing steps	Imaging data postprocessing, molecular data postprocessing, pathology data postprocessing
Use cases	Nephroblastoma
Scale (tissue, organ, cell etc.)	Multiscale heterogeneous data (clinical, imaging, molecular, pathology)
Challenges/Benefits/Limitations, Input Data standardisation	Poor standardization of DICOM data from different MRIs using different protocols, postprocessing of imaging data is needed to render the tumour, no automatic tools are available, combining different hypomodels to one hypermodel.
Existing standards (formats, guidelines, ontologies)	no information available
Model validation	Will be done by comparing the predictions with reality (imaging data after neoadjuvant chemotherapy and pathology data after surgery)
Lessons and Comments	This is done in close collaboration with Prof. Dr. Stamatakos and his group from ICCS, National Technical University of Athens, Greece. This shows that multidisciplinary approaches and team work are needed, including a legal and ethical framework. Sustainability is of importance after funding periods. The Medical Device Regulation needs to be taken into consideration if the model is to be used in clinical care.

Project	**iPlacenta: Integrative placenta: A systems biology approach towards phenotype-specific interactomes for placental function (https://www.iplacenta.eu/)**
Methods	a) Molecular interaction map construction and analysis; supervised ML b) In vitro cell work: expression of specific proteins/modified proteins in cells, differentiation of pluripotent stem cells into trophoblasts c) Use of small animal model for assessment of vasculature and endothelial function with the newly developed devices d) Doppler ultrasound in pregnancy women and collection of blood sample after the delivery. e) Molecular laboratory techniques, data analysis, bioinformatics f) Expression-Quantitative Trait Loci analysis by combining genotype and gene expression datasets from two cohorts of human placental samples. g) Recruitment patients and performing non-invasive CV assessment; supervised BT and AK
Model function	a) Identification of regulatory motifs; risk prediction b) Uncover mechanistic roles of redox modifications in angiogenic signalling and assess their functional effects in early pregnancy events c) Literature, in vivo data animals, knowledge industrial partner d) Identification of a epigenetic marker in maternal blood e) identification of senescence markers and disease progress f) Identification of eQTLs, of regulatory hot-spots g) Identification of abnormal CV findings; risk prediction
Input data	a) Primary Literature, public databases; clinical data (hospitalomics) b) Gene and protein expression, functional assays, proteomics, splicing micro-array c) Prototypes, in vivo validated d) Primary Literature, clinical data (from patients) e) Clinical and experimental data f) Genotype data (Acquired with InfiniumOmniexpress Illumina Array, from DNA samples) and gene expression data (Acquired with ClariomD microarray from Affymetrix, from RNA samples) g) Clinical data, biophysical and biochemical data
Output	a) Therapeutic drug target, biomarker, improved micro-arrays, knowledgebase; patient stratification & classification b) New insights on signaling pathways and cell functions c) Prototype development process d) Possibility of predicting postnatal neurological damage when the fetus is in utero. e) Understanding pathophysiology of adverse pregnacies affected by placental ageing. Biomarker identification and Senolytic therapeutics f) list of gene-snps that show statistically significant correlation and could indicate potential regulatory mechanisms in the placenta g) Patient stratification & classification; prediction; improving CV health in women
Processing steps	a) Various approaches b) --- c) Endothelial function of rodents, placenta vasculature in murine pregnancy

	d) Doppler ultrasound/Collection of blood samples - Analysis of samples/Analysis of results e) Various approaches f) Filtering and quality control of genotype and gene expression data g) Various approaches
Use cases	a) Pregnancy complications, here: preeclampsia intrauterine growth restriction b) Pathologies involving oxidative stress and angiogenesis, here preeclampsia c) Organ d) Pregnancy complications: late onset intrauterine growth restriction e) Pregnancy complications, here: preeclampsia intrauterine growth restriction f) Placental function pregnancy complications, here: preeclampsia intrauterine growth restriction g) Hypertensive disorders of pregnancy
Scale (tissue, organ, cell etc.)	a) Multiscale from placenta at the highest level to molecule on the lowest level; b) Molecules (single protein functions), cell (gene expression and functional effects, potentially interaction between several cell types in co-cultures) c) Animal study under home office license Comparison with other developments (similar technique) d) Multiscale (blood sample, extraction of micro-RNA) e) Multiscale from placenta at the highest level to molecule on the lowest level; f) Using DNA and RNA from the whole placenta (organ) g) Assessment of CV system and maternal heart
Challenges/Benefits/Limitations, Input Data standardisation	a) Dependency on results of partner projects results; Ultrasound image anonymization; hospital database access (law restrictions) b) Challenges/limitations: translation from molecular to tissue/organ scale, lack of physiological relevance for immortalised cell lines, questions towards precise identity of iPSC-derived trophoblasts; Benefits: precise focus on specific molecular pathways, easy access to and maintenance of cells c) --- d) Limitations: -Recruitment of patients and blood samples, follow-up of patients. -Benefits: possibility of predicting postnatal neurological damage when the fetus is in utero. -anonymization of the clinical data e) Sample collection and sample size; limited access to hospital data f) Main limitations are due to the small number of samples available. g) Recruitment and follow patients up/offering extra-care for patients/observational study
Existing standards (formats, guidelines, ontologies)	a) SBGN PD & AF (CellDesigner mix), GO, ChEBI, UniProt b) Published literature c) --- d) --- e) SBGN PD & AF (CellDesigner mix), GO, ChEBI, UniProt

	f) Study design based on previous studies found in the literature, integrating different approaches to better tackle our own dataset g) Not applicable
Model validation	a) Experimental validation, expert curation/validation b) Experimental in vitro validation with complementary studies, animal studies c) --- d) Experimental validation e) Experimental validation, expert curation/validation f) Overlap of findings with previous eQTL analyses in placenta, experimental validation g) External validation
Lessons and Comments	a) iPlacenta is an interdisciplinary project/training network that investigates pregnancy complications from with various approaches. Thus, the methods, data, standards and validation depend heavily on the sub project. Here we only list the information for b) --- c) --- d) iPlacenta is an interdisciplinary project/training network that investigates pregnancy complications from with various approaches. Thus, the methods, data, standards and validation depend heavily on the sub project. e) iPlacenta is an interdisciplinary project/training network that investigates pregnancy complications from with various approaches. Thus, the methods, data, standards and validation depend heavily on the sub project. f) --- g) iPlacenta is a great network among ESRs and among research teams all around Europe. I am very happy to be part of this.

Project	**LifeCycle: EU child cohort network (https://lifecycle-project.eu/)**
Methods	Regression analyses; Omic studies
Model function	no information available
Input data	Data collected in ongoing population-based cohorts studies in pregnancy and childhood (questionnaire, physical examinations, biomarkers, imaging, omics)
Output	Harmonized data for core exposure, covariate and outcome variables
Processing steps	Different approaches
Use cases	no information available
Scale (tissue, organ, cell etc.)	Mostly blood biomarkers and omics
Challenges/Benefits/Limitations, Input Data standardisation	Harmonisation steps to align 20 cohorts with over 300,000 participants is a challenge; lack of governance structure for data sharing through an federated data analysis approach
Existing standards (formats, guidelines, ontologies)	Not really available, we had to develop
Model validation	no information available
Lessons and Comments	harmonized data crucial for cross cohort collaboration; great opportunity to capitalize on existing data; GDPR complicates international collaboration

Project	**Multiple MS: Multiple manifestations of genetic and non-genetic factors in Multiple Sclerosis disentangled with a multi-omics approach to accelerate personalised medicine (https://www.multiplems.eu/)**
Methods	Unsupervised approach: e.g. Topological mapping, multi-partite "knowledge graph". Supervised approaches: e.g. cell-specific pathway analysis coupled to burden score.
Model function	Stratification of patients with MS
Input data	Genetic, lifestyle (questionnaires), established biomarkers. We will have two complementary approaches one using risk factors for MS and one using risk factors for disease severity measured in several different ways. For smaller part of cohort expression and methylome data will also be used.
Output	Clusters of patients that we then will be characterized clinically (i.e. Do the different clusters differ with regard to response to treatment or severity of disease?). Clinical data is already collected.
Processing steps	Data has been collected from several previous studies. Harmonization of data (both genetic, questionnaire and clinical).
Use cases	no information available
Scale (tissue, organ, cell etc.)	Genotyping done on blood. Biomarker analysis on blood or CSF. MRI of brain and spinal cord.
Challenges/Benefits/Limitations, Input Data standardisation	Harmonization of data from >30 previous studies have been challenging. Currently genetic, clinical, biomarker and MRI data have been harmonized. Lifestyle exposures have not been harmonized yet.
Existing standards (formats, guidelines, ontologies)	ICD10 used for comorbidities. Standard formats used for genotype data. Biomarker data as much as possible standardized to units used in clinical medicine. MRI, standardized pipeline used for processing DICOM images, this standard was developed in this project, but applied in several other project too.
Model validation	Models that are developed using data in the retrospective arm of the project will be validated in the prospective observational trial of newly diagnosed MS patients which is also part of the study.
Lessons and Comments	We have learnt that harmonization of data takes longer than we expected. We have also learnt that getting data processing agreements in place allowing sharing of data was complicated and took a lot of effort but is not impossible. We are only now starting the modelling part of the project.

Project	**Personalised treatment of anaemia in lung cancer patients (https://www.dkfz.de/en/systembiologie/AreasofInterest.html)**
Methods	no information available
Model function	Mathematical model (ODE) predicting outcome of treatment options based on hemoglobin and CRP values (longitudinal measurements and cohort data).
Input data	Lab values (Hgb, CRP)
Output	Treatment outcome predictions
Processing steps	no information available
Use cases	no information available
Scale (tissue, organ, cell etc.)	Modelling on the Epo pathway and coupling to whole body effects in anemia
Challenges/Benefits/Limitations, Input Data standardisation	Non-standardised input data, e.g. homemade medication name input data. Challenges: input data variation where input data was not recognized by the model as the same treatment under two different names. Challenges: importance of knowing precise death time and date including which day of the week; perceived conflict with principles of data minimization. Challenge: does the model take into account the subjective well-being of the patient, as an outcome?
Existing standards (formats, guidelines, ontologies)	no information available
Model validation	Model validation: Comparison of patient outcome with or without use of the model, e.g. survival. Design? RCT parallel populations? Difficult to test the model in different hospital because of homemade input data standards, but not impossible.
Lessons and Comments	no information available

Project	**SYSCID: Systems medicine for chronic inflammatory diseases (https://syscid.eu/)**
Methods	Machine learning. Analytical pipelines including but not limited to standard methods DNA: BWA/Samtools/GATK RNA: STAR/DeSEQ2 DNA methylome: Bismark/RnBeads Microbiome: Mothur/Qiime/Humman2/Metaphlan Marker identification and reduction to diagnostic sets are done by different ML approaches (random forest , Bayesian inference)
Model function	Biomarker identification; predict disease outcome and treatment response to guide therapy decisions on individual basis.
Input data	Tissue and blood coupled to clinical data from EHR and PRO. Exomes/genomes, transcriptomes, DNA methylomes and 16rRNA /metagenomics (microbiome) data.
Output	Several data and metadata formats one all analysed level according to standards (DNA/DNAm/RNA) according to IHEC guidelines.
Processing steps	Oriented towards research question, the above mentioned pipelines use standard data processing (e.g. Deseq2).
Use cases	Inflammatory diseases (IBD, RA and SLE)
Scale (tissue, organ, cell etc.)	Tissue, blood and single cell from peripheral leukocytes
Challenges/Benefits/Limitations, Input Data standardisation	Versioning of community standards (e.g. reference genomes and updates of mappers and count software)
Existing standards (formats, guidelines, ontologies)	Single cell field developing quickly, less standardized compared to genetic analyses. As IHEC and HMP (Raes) Partners SYSCID is well aware of data standards. For response analyses, longitudinal analyses and models building on regulatory /functional networks are necessary. Here, we feel that there is much less standardization. This is an unmet need in the systems immunology field.
Model validation	no information available
Lessons and Comments	Importance of standardizing outcome data that is meaningful for patients. Increase communication between modellers and medical specialization communities, patient organisations.

References

Akanksha Limaye, D. A. N. (2019). Machine learning models to predict the precise progression of Tay-Sachs and Related Disease. MOL2NET 2019, International Conference on Multidisciplinary Sciences, 5th edition session USEDAT-07: USA-Europe Data Analysis Training School, UPV/EHU. Bilbao-JSU, Jackson, USA, 2019

Bender, D. and K. Sartipi (2013). "HL7 FHIR: An Agile and RESTful Approach to Healthcare Information Exchange." 2013 Ieee 26th International Symposium on Computer-Based Medical Systems (Cbms): 326-331.

Bergmann, F. T., R. Adams, S. Moodie, J. Cooper, M. Glont, M. Golebiewski, M. Hucka, C. Laibe, A. K. Miller, D. P. Nickerson, B. G. Olivier, N. Rodriguez, H. M. Sauro, M. Scharm, S. Soiland-Reyes, D. Waltemath, F. Yvon and N. Le Novere (2014). "COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project." BMC Bioinformatics **15**: 369.

DICOM_Secretariat. (2020). "Digital Imaging and Communications in Medicine ", from https://www.dicomstandard.org/.

GA4GH_Community. (2020). "The Global Alliance for Genomics and Health." from https://www.ga4gh.org/.

Georgiadi, E. C., D. D. Dionysiou, N. Graf and G. S. Stamatakos (2012). "Towards in silico oncology: Adapting a four dimensional nephroblastoma treatment model to a clinical trial case based on multi-method sensitivity analysis." Computers in Biology and Medicine **42**(11): 1064-1078.

GitHub_Community. (2020). "GitHub." from https://github.com/features.

Goddard, N. H., M. Hucka, F. Howell, H. Cornelis, K. Shankar and D. Beeman (2001). "Towards NeuroML: Model description methods for collaborative modelling in neuroscience." Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **356**(1412): 1209-1228.

Golebiewski, M. (2019). Data Formats for Systems Biology and Quantitative Modeling. Encyclopedia of Bioinformatics and Computational Biology**:** 884-893.

Griss, J., A. R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G. G. Thallinger, R. M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, S. Neumann, J. Fan, F. Reisinger, Q. W. Xu, N. Del Toro, Y. Perez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaino and H. Hermjakob (2014). "The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience." Mol Cell Proteomics **13**(10): 2765-2775.

Hsi-Yang Fritz, M., R. Leinonen, G. Cochrane and E. Birney (2011). "Efficient storage of high throughput DNA sequencing data using reference-based compression." Genome Research **21**(5): 734-740.

Hucka, M., F. T. Bergmann, A. Drager, S. Hoops, S. M. Keating, N. Le Novere, C. J. Myers, B. G. Olivier, S. Sahle, J. C. Schaff, L. P. Smith, D. Waltemath and D. J. Wilkinson (2018). "The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core." J Integr Bioinform **15**(1): 1-173.

Kalra, D., T. Beale and S. Heard (2005). "The openEHR Foundation." Stud Health Technol Inform **115**: 153-173.

Kemp, B., A. Varri, A. C. Rosa, K. D. Nielsen and J. Gade (1992). "A simple format for exchange of digitized polygraphic recordings." Electroencephalogr Clin Neurophysiol **82**(5): 391-393.

Kuepfer, L., C. Niederalt, T. Wendl, J. F. Schlender, S. Willmann, J. Lippert, M. Block, T. Eissing and D. Teutonico (2016). "Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model." CPT Pharmacometrics Syst Pharmacol **5**(10): 516-531.

Le Novere, N., A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence and B. L. Wanner (2005). "Minimum information requested in the annotation of biochemical models (MIRIAM)." Nat Biotechnol **23**(12): 1509-1515.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Lipman, D. and W. Pearson (1985). "Rapid and sensitive protein similarity searches." Science **227**(4693): 1435-1441.

Lloyd, C. M., M. D. Halstead and P. F. Nielsen (2004). "CellML: its future, present and past." Prog Biophys Mol Biol **85**(2-3): 433-450.
Madsen, C., A. G. Moreno, P. Umesh, Z. Palchick, N. Roehner, C. Atallah, B. Bartley, K. Choi, R. S. Cox, T. Corochowski, R. K. Grunberg, C. Macklin, J. McLaughlin, X. W. Meng, T. Nguyen, M. Pocock, M. Samineni, J. Scott-Brown, Y. Tarter, M. Zhang, Z. Zhang, Z. Zundel, J. Beal, M. Bissell, K. Clancy, J. H. Gennari, G. Misirli, C. Myers, E. Oberortner, H. Sauro and A. Wipat (2019). "Synthetic Biology Open Language (SBOL) Version 2.3." Journal of Integrative Bioinformatics **16**(2).
Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz and E. W. Deutsch (2011). "mzML--a community standard for mass spectrometry data." Mol Cell Proteomics **10**(1): R110 000133.
Neal, M. L., M. Konig, D. Nickerson, G. Misirli, R. Kalbasi, A. Drager, K. Atalag, V. Chelliah, M. T. Cooling, D. L. Cook, S. Crook, M. de Alba, S. H. Friedman, A. Garny, J. H. Gennari, P. Gleeson, M. Golebiewski, M. Hucka, N. Juty, C. Myers, B. G. Olivier, H. M. Sauro, M. Scharm, J. L. Snoep, V. Toure, A. Wipat, O. Wolkenhauer and D. Waltemath (2019). "Harmonizing semantic annotations for computational models in biology." Brief Bioinform **20**(2): 540-550.
Nickerson, D., K. Atalag, B. de Bono, J. Geiger, C. Goble, S. Hollmann, J. Lonien, W. Muller, B. Regierer, N. J. Stanford, M. Golebiewski and P. Hunter (2016). "The Human Physiome: how standards, software and innovative service infrastructures are providing the building blocks to make it achievable." Interface Focus **6**(2): 20150103.
Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proceedings of the National Academy of Sciences **85**(8): 2444-2448.
Rosse, C. and J. L. V. Mejino (2003). "A reference ontology for biomedical informatics: the Foundational Model of Anatomy." Journal of Biomedical Informatics **36**(6): 478-500.
Schreiber, F., G. D. Bader, P. Gleeson, M. Golebiewski, M. Hucka, N. Le Novere, C. Myers, D. Nickerson, B. Sommer and D. Waltemath (2016). "Specifications of Standards in Systems and Synthetic Biology: Status and Developments in 2016." Journal of Integrative Bioinformatics **13**(3).
Schreiber, F., B. Sommer, G. D. Bader, P. Gleeson, M. Golebiewski, M. Hucka, S. M. Keating, M. Konig, C. Myers, D. Nickerson and D. Waltemath (2019). "Specifications of Standards in Systems and Synthetic Biology: Status and Developments in 2019." J Integr Bioinform **16**(2): 1-5.
Shefchek, K. A., N. L. Harris, M. Gargano, N. Matentzoglu, D. Unni, M. Brush, D. Keith, T. Conlin, N. Vasilevsky, X. A. Zhang, J. P. Balhoff, L. Babb, S. M. Bello, H. Blau, Y. Bradford, S. Carbon, L. Carmody, L. E. Chan, V. Cipriani, A. Cuzick, M. D. Rocca, N. Dunn, S. Essaid, P. Fey, C. Grove, J. P. Gourdine, A. Hamosh, M. Harris, I. Helbig, M. Hoatlin, M. Joachimiak, S. Jupp, K. B. Lett, S. E. Lewis, C. McNamara, Z. M. Pendlington, C. Pilgrim, T. Putman, V. Ravanmehr, J. Reese, E. Riggs, S. Robb, P. Roncaglia, J. Seager, E. Segerdell, M. Similuk, A. L. Storm, C. Thaxon, A. Thessen, J. O. B. Jacobsen, J. A. McMurry, T. Groza, S. Kohler, D. Smedley, P. N. Robinson, C. J. Mungall, M. A. Haendel, M. C. Munoz-Torres and D. Osumi-Sutherland (2020). "The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species." Nucleic Acids Res **48**(D1): D704-D715.
Stamatakos, G., D. Dionysiou, A. Lunzer, R. Belleman, E. Kolokotroni, E. Georgiadi, M. Erdt, J. Pukacki, S. Rueping, S. Giatili, A. d' Onofrio, S. Sfakianakis, K. Marias, C. Desmedt, M. Tsiknakis and N. Graf (2014). "The Technologically Integrated Oncosimulator: Combining Multiscale Cancer Modeling With Information Technology in the In Silico Oncology Context." Ieee Journal of Biomedical and Health Informatics **18**(3): 840-854.
Waltemath, D., R. Adams, F. T. Bergmann, M. Hucka, F. Kolpakov, A. K. Miller, Moraru, II, D. Nickerson, S. Sahle, J. L. Snoep and N. Le Novere (2011). "Reproducible computational biology experiments with SED-ML--the Simulation Experiment Description Markup Language." BMC Syst Biol **5**: 198.

Acknowledgements

EU-STANDS4PM is funded by the European Union Horizon2020 framework programme of the European Commission Directorate-General for Research and Innovation under Grant Agreement # 825843.

The current report is part of WP1 Data sources and standards for predictions in personalized medicine, and Deliverable 1.2 *EU-wide mapping report on good practice examples for integrating phenotype and large scale data.*

Contact

Martin Golebiewski
HITS gGmbH Heidelberg
Germany
E-Mail: martin.golebiewski@h-its.org

Marc Kirschner (Coordination)
Forschungszentrum Jülich GmbH
Project Management Jülich (PtJ)
52425 Jülich, Germany
Phone: +49 2461 61-6863
E-mail: m.kirschner@fz-juelich.de

Sylvia Krobitsch
Forschungszentrum Jülich GmbH
Project Management Jülich (PtJ)
52425 Jülich, Germany
Phone: + 49 30 20199-3403
E-mail: s.krobitsch@fz-juelich.de

www.eu-stands4pm.eu


EU-STANDS4PM
standards for in silico models
for personalised medicine