EU-STANDS4PM
standards for in silico models for personalised medicine

A European standardization framework for data integration and data-driven *in-silico* models for personalised medicine

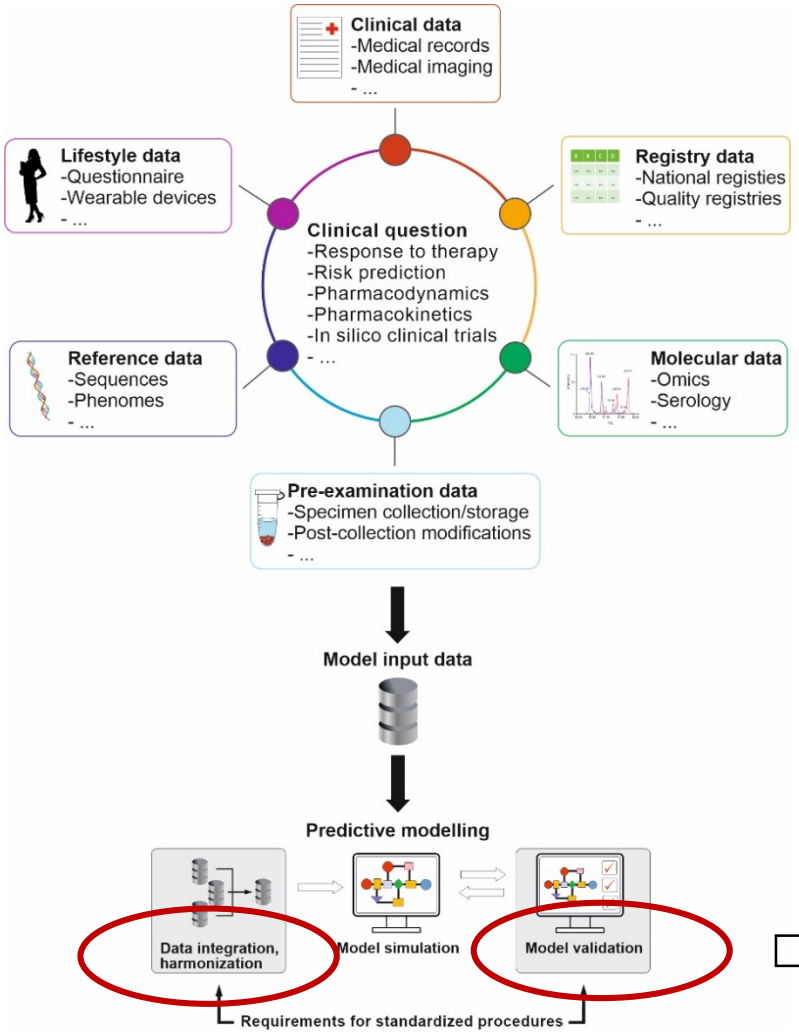
SimCardioTest workshop 08.02.2022

Marc Kirschner, coordinator on behalf of the EU-STANDS4PM consortium
m.kirschner@fz-juelich.de

EU-STANDS4PM at a glance

- **Type:** Coordination and Support Action
- **European Commission:** H2020 Work Programme 2018-2020
- **Project duration:** 3,5 years (Jan 2019- Jun 2022)

Computational models for personalized medicine – data integration and model validation



Recommendations and standards for

- ⇨ Data integration
- ⇨ Model validation
- ⇨ Legal/ethical issues (e.g. patient rights, GDPR)

Target communities

- ⇨ European collaborative research
- ⇨ Funding organizations

Key objective: Development of formal standardization documents
ISO-Technical Specification for computational modelling approaches

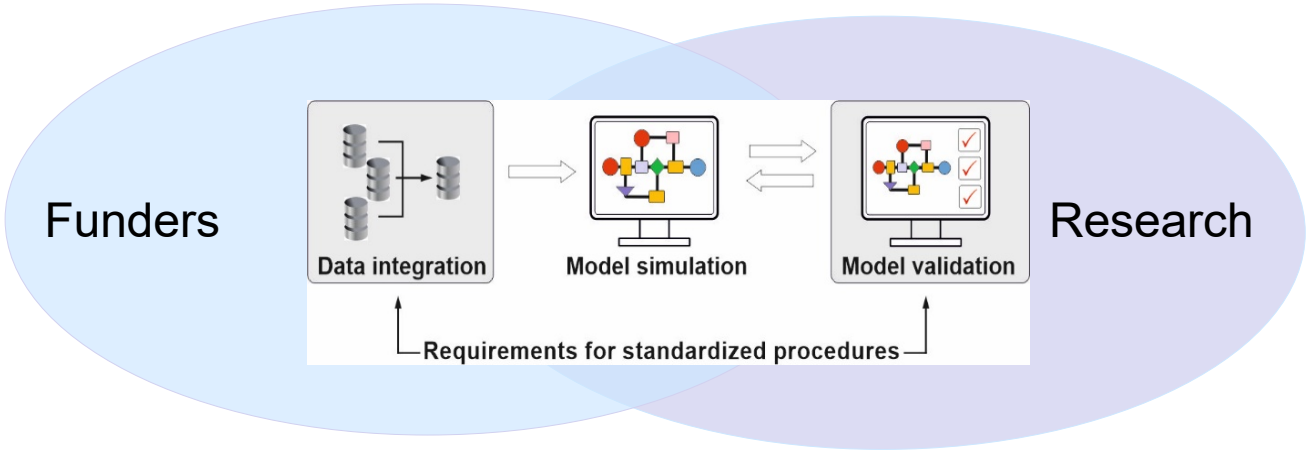
Rationale to develop an ISO Technical Specification

Currently

- ⇨ No broadly accepted standards for health & disease data integration for predictive modelling
- ⇨ Many grass root but non-binding standards
- ⇨ Fragmentation

Clear need for common standards for

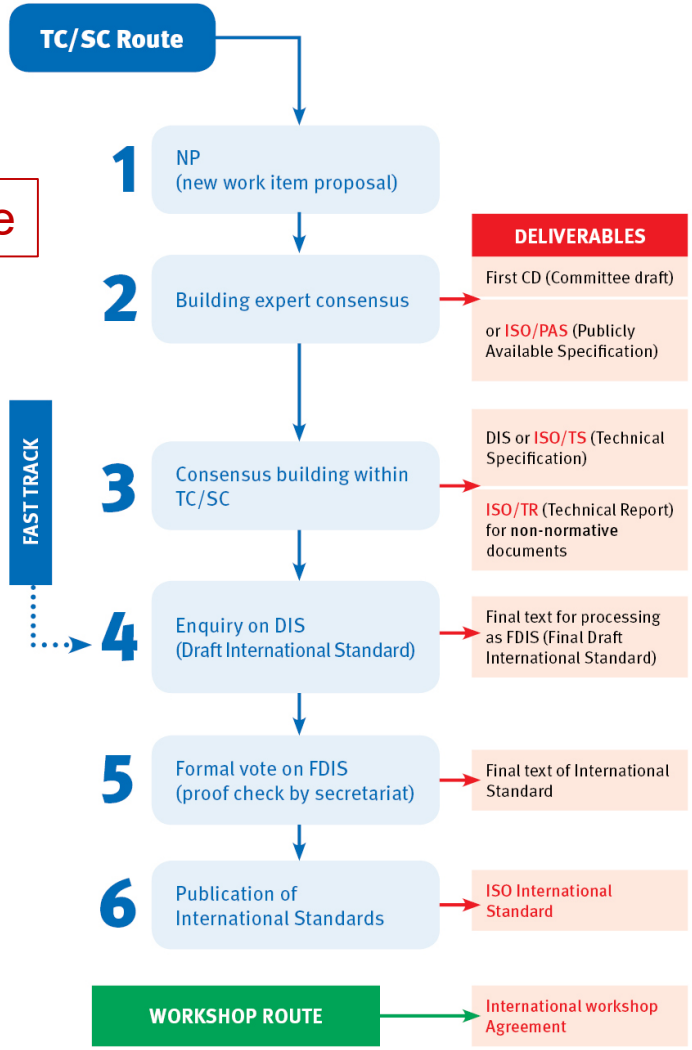
- ⇨ for data integration and
- ⇨ model validation
- ⇨ a broad adaptation of theses standards in health and clinical research



ISO Deliverables

- ISO Standards
- ISO/TS Technical Specifications
- ISO/TR Technical Reports
- ISO/PAS Publicly Available Specifications
- IWA International Workshop Agreements
- ISO Guides

normative



Source: www.iso.org

ISO Technical Specification

A Technical Specification addresses **work still under technical development**, or where it is believed that **there will be a future**, but not immediate, possibility of agreement on an International Standard. A Technical Specification is **published for immediate use**, but it also provides a means to obtain feedback. The aim is that it **will eventually be transformed and republished as an International Standard**.

Source: www.iso.org

ISO-Technical Specification for computational modelling approaches

ISO/AWI TS 9491

Biotechnology — Recommendations and requirements for predictive computational models in personalised medicine research — Part 1: Guidelines for constructing, verifying and validating models

Scope ISO/AWI TS 9491

This document defines challenges and requirements for predictive **computational models constructed for research** purposes in personalised medicine. It specifies **recommendations and requirements** for the setup, formatting, validation, simulation, storing and sharing of such models, as well as their application in clinical trials and other research areas. It summarizes specific **challenges regarding data input, as well as verifying and validating of such models** that can be considered as best practices for modelling in research and development in the field of personalised medicine.

This document also specifies recommendations and requirements for data used to construct or needed for validating models, including rules and requirements for formatting, description, annotation, interoperability, integration, accessing, as well as recording and documenting the provenance of such data.

This document **does not** provide specific rules or requirements for the use of computational models in the **clinical routine, or for diagnostic or therapeutic purposes.**

Contents

Content example: Models

5.2.2 Cellular systems biology models

Purpose: Simulation of complex dynamic biological processes and networks. Models can be either data-driven ("bottom-up") or mechanism-based ("top-down").
Mechanism-based concepts aim for a structural representation of the governing physiological processes based on model equations with limited amount of data, which are required for the base model establishment [16] or, alternatively, on static interacting networks [17, 18]. Data-driven approaches [4, 19] require, as the name implies, sufficiently rich and quantitative time-course data to train and to validate the model. Due to its often black-box nature, the model validation process in data-driven approaches relies on performance tests against known results.

Challenges

- Creation of models that balance the level of abstraction with comprehensiveness to make modelling efforts reproducible and reusable (abstraction vs size).
- Development of prediction models that can be adopted easily to individual patient profiles.
- Efficient parameter estimation tools to cope with population and disease heterogeneity.
- Overfitting of the model to the experimental/patient data and optimization methods for model predictions in a realistic parametric uncertainty.
- Flexibility in models to cope with missing data (e.g., diverse patient profiles).
- Scaling from cellular to organ and to organism levels (e.g., high clinical relevance, high hurdles for regulatory acceptancy).

5.2.3 Risk prediction for common diseases

Purpose: Prediction of the risk of an individual having a disease based on the individual's genetic make-up (Polygenetic Risk Score, PRS)[20-22]. Polygenic models assume that the genetic variance of a disease is a combination of small effects of multiple variants across the allele frequency spectrum. Genome-wide association studies (GWAS) that scan the genomes of thousands of individuals offer a very powerful method to identify these multiple genetic risk factors for having the disease.

Challenges

- Understanding the possible implication to patients at individual-level, what can be inferred? How to test the inference made?
- Limited replication of genetic associations and poor application of diverse populations (e.g., too poorly represented to be of interest for specific analyses), specifically of mixed or non-European ancestry.
- Varying transparency of methodological choices and reproducibility.
- Limited cellular/tissue context and harmonized functional data availability across populations/studies.
- Missing environmental information coupled to genetic data.

Content example: Data preparation

5.4 Data preparation for integration into computer models

5.4.1 General

Computational models in the life sciences in general and in personalised medicine research specifically are increasingly incorporating rich and varied datasets to capture multiple aspects of the modelled phenomenon. Data types are encoded in technology and subdomain specific formats and the variety and incompatibility, as well as lack of interoperability of such data formats have been noted as one of the major hurdles for data preparation.

To allow for seamless integration of data used for the construction of predictive computational models in personalised medicine these data shall:

- Include or be annotated with sampling and specimen data that follow the recommendations and requirements as specified by the relevant domain-specific standards
- Be formatted using generally accepted and interoperable standard data formats commonly used for the corresponding data types (as specified by ISO 20691)
- Include or be annotated with descriptive metadata that consider generally accepted domain-specific Minimum Information guidelines and describes the metadata attributes and entities using semantic standards, such as standard terminologies, controlled vocabularies and ontologies (as specified in annex B of ISO 20691)
- Follow best practice recommendations and requirements of generally accepted domain-specific data interoperability frameworks
- Structured in a way that allows integration of the data into a model, also together with other data
- Include or be annotated with data provenance information that allows for tracking of the data and source material throughout the whole data processing and modelling
- Be made accessible via harmonized Data Access Agreements (hDAAs) for controlled access data, if open access to the data is not possible

5.4.2 Sampling data

Dedicated measures need to be taken for collecting, stabilizing, transporting, storing and processing of biological specimen/samples, to ensure that profiles of analytes of interest (e.g. gene sequence, transcript, protein, metabolite) for examination are not changed ex vivo. Without these measures, analyte profiles can change drastically during and after specimen collection, thus making the outcome from diagnostics or research unreliable or even impossible, because the subsequent examination will not determine the situation in the patient, but an artificial profile generated during the pre-examination process. Important measures include for example times and temperatures of sample transportation not exceeding the specifications provided in ISO standards and CEN technical specifications (e.g. ISO 20658, ISO 20916, ISO 20186), giving guidelines on all steps of the pre-examination workflow. Conditions applied to specimen are documented in addition to other important metadata, including but not limited to the following Table 2.

Content example: Model validation

5.6 Model validation

5.6.1 General

Common to all in silico models is a need for validation [53] and accuracy, however, in contrast to data input model validation methods are considered to be individual and type-specific (Table 4). It is important that any algorithm performs well on novel data that have not been used in training the algorithm; i.e. the model should be able to generalize to new data from the same domain [54]. There are guidelines and methods for validating models, which are accurate and confident in predictions, both in terms of accuracy and confidence in predictions. Performance evaluation of in silico models should be transparent and consistent to existing guidelines, explaining the reasons for not doing so when alternative methods are chosen.

5.6.2 Specific recommendations for model validation

The following Table 4 contains a summary of type specific recommendations for different modelling approaches.

Table 4 — Specific recommendations for model validation

Cellular systems biology models
R1: Develop a standardized protocol for patient history and information and integrate with clinical standard protocols, e.g., electronic health records (EHRs) and FHIR. R2: Agree on the use of patient expression profile e.g., mRNA or protein concentration. R3: Develop prediction models with a fitting hypothesis. R4: Use model replication and reproduction before considering clinical trials. R5: Compare model predictions (e.g., biomarkers selection) with established clinical ones. R6: Develop of user-friendly graphical interfaces to ease the use of models in clinic.
Risk prediction for common diseases
R1: Ensure highest possible diversity and sample sizes further for all the genetic studies on complex diseases as well as for performing functional studies. R2: Enable more transparent, standardized and detailed methods clearly stating methodological choices made with necessary justifications to enhance reproducible research.
Disease course and therapy response prediction
R1: Harmonize disease specific scores including objective parameters of disease activity and progression. R2: Support the development and validation of innovative patient-reported outcome tools for clinical trials including the safe and easy use of app- and wearable-based technologies. R3: Define minimum clinical criteria for a systems medicine trial combined using harmonized scores and quantitative, validated patient reported outcomes into account. R4: Develop concepts for integrating of model-based prediction and AI content into curricular education in medicine and medical life sciences. R5: Commence stakeholder discussions including political decision makers to overcome financial and intellectual hurdles in large European consortia trials, ideally involving both academic and European industry partners (current schemes of the Innovative Medicine Initiative[24] could serve as a starting point). R6: Develop new models of public-private partnerships for a strong European health care economy on IP

The players and their key tasks

H2020 core projects/ data governance pilot

UCL — Stephan Beck

Università di Parma, BD2Decide — Tito Poli

QIAGEN, SPIDIA4P — Uwe Ölmüller

CAU Christian-Albrechts-Universität zu Kiel, SYSCID — Philip Rosenstiel

Rotterdam University, LifeCycle — Vincent Jaddoe

Karolinska Institutet, MultipleMS — Ingrid Kockum

Standardization

DIN — Heike Moser

HITS — Martin Golebiewski

QIAGEN — Kalle Günther

Data and models

elixir — Niklas Blomberg

University of Copenhagen — Catherine Collin

Universität Rostock — Olaf Wolkenhauer

RWTH Aachen University — Lars Küpfer

Legal/ethical frame

University of Copenhagen — Mette Hartlev

University of Oxford — Miranda Mourby

Vilniaus Universitetas 1579 — Eugenijus Gefenas

Regulatory support

famhp federal agency for medicines and health products — Flora M. Tschinanu

Coordination

JÜLICH

ptj Projektträger Jülich Forschungszentrum Jülich

Marc Kirschner
Sylvia Krobitsch

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.