TECHNICAL SPECIFICATION

ISO/TS 9491-1

First edition
2023-06

Biotechnology — Predictive computational models in personalized medicine research —

Part 1: Constructing, verifying and validating models

Biotechnologie — Modèles informatiques prédictifs dans la recherche sur la médecine personnalisée —

Partie 1: Construction, vérification et validation des modèles

Reference number
ISO/TS 9491-1:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 276, *Biotechnology*.

A list of all parts in the ISO 9491 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

The capacity to generate data in life sciences and health research has greatly increased in the last decade. In combination with patient/personal-derived data, such as electronic health records, patient registries and databases, as well as lifestyle information, this big data holds an immense potential for clinical applications, especially for computer-based models with predictive capacities in personalized medicine. However, and despite the ever-progressing technological advances in producing data, the exploitation of big data to generate new knowledge for medical benefits, while guaranteeing data privacy and security, is lacking behind its full potential. A reason for this obstacle is the inherent heterogeneity of big data and the lack of broadly accepted standards allowing interoperable integration of heterogeneous health data to perform analysis and interpretation for predictive modelling approaches in health research, such as personalized medicine.

Common standards lead to a mutual understanding and improve information exchange within and across research communities and are indispensable for collaborative work. In order to setup computer models in personalized medicine, data integration from heterogeneous and different sources at different times plays a key role. Consistent documentation of data, models and simulation results based on basic guiding principles for data management practices, such as FAIR (findable, accessible, interoperable, reusable)[7] or ALCOA (attributable, legible, contemporaneous, original, accurate), and standards can ensure that the data and the corresponding metadata (data describing the data and its context), as well as the models, methods and visualizations, are of reliable high quality.

Hence, standards for biomedical and clinical data, simulation models and data exchange are a prerequisite for reliable integration of health-related data. Such standards, together with harmonized ways to describe their metadata, ensure the interoperability of tools used for data integration and modelling, as well as the reproducibility of the simulation results. In this sense, modelling standards are agreed ways of consistently structuring, describing, and associating models and data, their respective parts and their graphical visualization, as well as the information about applied methods and the outcome of model simulations. Such standards also assist in describing how constituent parts interact, or are linked together, and how they are embedded in their physiological context.

Major challenges in the field of personalized medicine are to:

a) harmonize the standardization efforts that refer to different data types, approaches and technologies;

b) make the standards interoperable, so that the data can be compared and integrated into models.

An overall goal is to FAIRify data and processes in order to improve data integration and reuse. An additional challenge is to ensure a legal and ethical framework enabling interoperability.

This document presents modelling requirements and recommendations for research in the field of personalized medicine, especially with focus on collaborative research, such that health-related data can be optimally used for translational research and personalized medicine worldwide.

Biotechnology — Predictive computational models in personalized medicine research —

Part 1: Constructing, verifying and validating models

1 Scope

This document specifies requirements and recommendations for the design, development and establishment of predictive computational models for research purposes in the field of personalized medicine. It addresses the set-up, formatting, validation, simulation, storing and sharing of computational models used for personalized medicine. Requirements and recommendations for data used to construct or required for validating such models are also addressed. This includes rules for formatting, descriptions, annotations, interoperability, integration, access and provenance of such data.

This document does not apply to computational models used for clinical, diagnostic or therapeutic purposes.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 20691:2022, *Biotechnology — Requirements for data formatting and description in the life sciences*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
artificial intelligence
AI
<system> capability to acquire, process, create and apply knowledge, held in the form of a model, to conduct one or more given tasks

[SOURCE: ISO/IEC TR 24030:2021, 3.1]

3.2
molecular biomarker
biomarker
molecular marker
detectable and/or quantifiable molecule or group of molecules used to indicate a biological condition, state, identity or characteristic or an organism

EXAMPLE Nucleic acid sequences, proteins, small molecules such as metabolites, other molecules such as lipids and polysaccharides.

[SOURCE: ISO 16577:2022, 3.4.28]

3.3
big data in health
high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points

[SOURCE: Reference [8]]

3.4
community standard
standard that reflects the results of a grass-roots standardization effort from a specific user group, and that is created by individual organizations or communities

3.5
computational model
in silico model
description of a system in a mathematical expression and/or graphical form highlighting objects and their interfaces

Note 1 to entry: An object distributed processing (ODP) concept.

Note 2 to entry: The computational model is similar to OMT ad UML notion of a class diagram when using the graphical form.

[SOURCE: ISO/IEC 16500-8:1999, 3.6, modified — Admitted term added. “mathematical expression and/or” added, and “as such it is similar to the OMT and UML notion of a class Diagram” deleted from the definition. “An object distributed processing (ODP) concept” moved to Note 1 to entry. Note 2 to entry added.]

3.6
data-driven model
model developed through the use of data derived from tests or from the output of investigated process

[SOURCE: ISO 15746-1:2015, 2.4]

3.7
data harmonization
technical process of bringing together different data types to make them processable in the same computational framework

3.8
data integration
systematic combining of data from different independent and potentially heterogeneous sources, to create a more compatible, unified view of these data for research purpose

[SOURCE: ISO 5127:2017, 3.1.11.24]

3.9
genome-wide association studies
GWAS
testing of genetic variants across the genomes of many individuals to identify genotype–phenotype associations

3.10
***in silico* clinical trial**
use of individualized computer simulation in the development or regulatory evaluation of a medicinal product, medical device or medical intervention

[SOURCE: Reference [9]]

3.11
***in silico* approach**
computer-executable analyses of *mathematical model(s)* (3.13) to study and simulate a biological system

3.12
machine learning
ML
computer technology with the ability to automatically learn and improve from experience without being explicitly programmed

EXAMPLE Speech recognition, predictive text, spam detection, artificial intelligence.

[SOURCE: ISO 20252:2019, 3.52, modified — Abbreviated term "ML" added.]

3.13
mathematical model
sets of equations that describes the behaviour of a physical system

[SOURCE: ISO 16730-1:2015, 3.11]

3.14
mechanism-based
approach in computational modelling that aims for a structural representation

3.15
model validation
comparison between the output of the calibrated model and the measured data, independent of the data set used for calibration

[SOURCE: ISO 14837-1:2005, 3.7]

3.16
model verification
confirmation that the mathematical elements of the model behave as intended

[SOURCE: ISO 14837-1:2005, 3.8]

3.17
personalized medicine
medical model using characterization of individuals' phenotypes and genotypes for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention

Note 1 to entry: Examples for individuals' phenotypes and genotypes are molecular profiling, medical imaging and lifestyle data.

Note 2 to entry: Medical decisions, prevention strategies and therapies in personalized medicine are based on this individuality.

[SOURCE: EU 2015/C 421/03[10]]

3.18
raw data
data in its originally acquired, direct form from its source before subsequent processing

[SOURCE: ISO 5127:2017, 3.1.10.04]

4 Principles

4.1 General

Research in the field of personalized medicine is highly dependent on the exchange of data from different sources, as well as harmonized integrative analysis of large-scale personalized medicine data (big data in health). Computational modelling approaches play a key role for understanding, simulating and predicting the molecular processes and pathways that characterize human biology. Modelling approaches in biomedical research also lead to a more profound understanding of the mechanisms and factors that drive disease, and consequently allow for adapting personalized treatment strategies that are guided by central clinical questions. Patients can greatly benefit from this development in research that equips personalized medicine with predictive capabilities to simulate *in silico* clinically relevant questions, such as the effect of therapies, the response to drug treatments or the progression of disease.

4.2 Computational models in personalized medicine

4.2.1 General

Computational models have the potential to translate *in vitro*, non-clinical and clinical results (and their related uncertainty) into descriptive or predictive expressions. The added value of such models in medicine and pharmacology has increasingly been recognized by the scientific community,[11][12][13][14] as well as by regulatory bodies such as the European Medicines Agency (e.g. EMA guideline on PBPK reporting[15]), or the US Food and Drug Administration (FDA).[16][17] Computational models are integrated in different fields in medicine and drug development expanding from disease modelling, molecular biomarker research to assessment of drug efficacy and safety. *In silico* approaches are also expanding in neighbouring fields, such as pharmacoeconomics,[18][19] analytical chemistry[20][21] and biology that are out of scope of this document.[22][23]

Model creation starts with a clinical question and the collection of data (see Figure 1). The data employed need harmonized approaches for data integration to start the model construction. The initial model usually undergoes several refinement and improvement iterations to enhance predictive capabilities. Common standards (see 4.3.3) should be used for the model building and curation process. Accuracy measurements and validation processes are key, and should be transparent, while model output and function should ideally be interpretable or explainable.

A number of computational modelling approaches in pre-clinical and clinical research already address these questions in detail (see 4.2.2 to 4.2.6) and, therefore, play a leading role for the future development of personalized medicine.

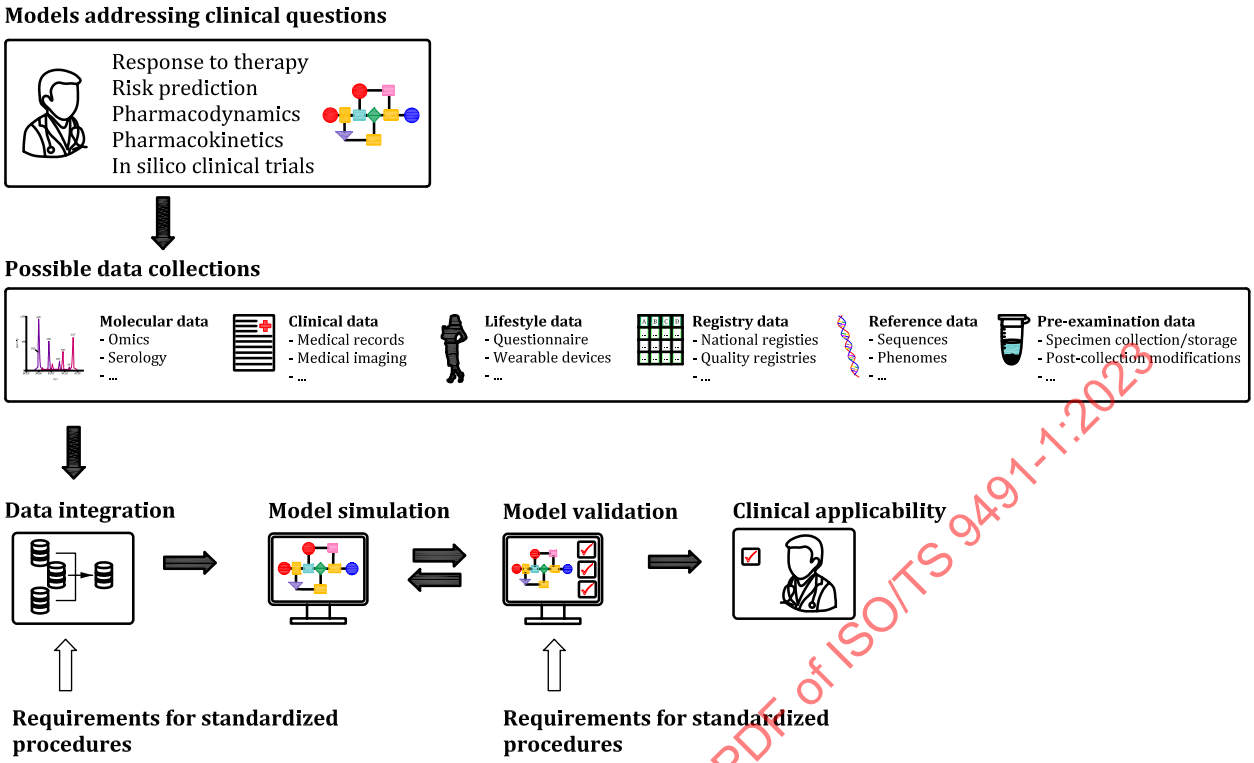

Figure 1 — Modelling approach for personalized medicine

4.2.2 Cellular systems biology models

4.2.2.1 General

For the 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[24] or, alternatively, on static interacting networks.[25][26] Data-driven approaches[11][27] require sufficiently rich and quantitative time-course data to train and to validate the model. Due to the often black-box nature of data driven approaches, the model validation process relies on performance tests against known results.

4.2.2.2 Challenges

The challenges are as follows:

— Creation of models that balance the level of abstraction with comprehensiveness to make modelling efforts reproducible and reusable (abstraction versus 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).

4.2.3 Risk prediction for common diseases

4.2.3.1 General

Predictive models stratify patients into distinct subgroups at different levels of risk for clinical outcomes (risk prediction for disease). By training the algorithm on clinical data, phenotypic or genotypic, subgroups can be identified which have identifiably different patterns of clinical markers. By then identifying which patterns a patient fits best, the model can place a particular patient within the most similar trajectory, thereby also stratifying the patient to a particular level of risk. Clinical markers used in such models can be any health feature, tokenized as to be analysable by the model, from data such as disease history symptoms, treatment and other exposure data, family history, laboratory data, etc., to genetic data.

4.2.3.2 Challenges

The challenges are as follows:

— Understanding the possible implication to patients at an 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.

4.2.4 Disease course and therapy response prediction

4.2.4.1 General

Prediction of the disease behaviour (mild versus severe, stable versus progressive) early in the disease course based on specific molecular biomarkers can allow an improved timing of therapy introduction, as well as the choice of therapy scheme (targeted therapy).[28] Ideally, these models can provide a prediction of multi-factorial diseases at unprecedented resolution, in a way that clinicians can use the information in their daily decision-making.

4.2.4.2 Challenges

The challenges are as follows:

— Harmonization and standardization of clinical information for measuring the disease of interest.

— Developing transparent and quality-controlled workflows for molecular data generation and interpretation in clinical settings.

— Harmonization and application of existing and upcoming pre-examination workflow standards (including specimen collection, storage and nucleic acid isolation), as well as developing feasible ring trial formats and external quality assurance (EQA) schemes for given molecular analysis types.

— Transparent reduction of contents and definition of appropriate marker sets and dynamic models to foster clinical translation.

— Developing intuitive visualization results and insights into molecular analyses, as well as critical appraisal of limitations of models by physicians.

4.2.5 Pharmacokinetic/-dynamic modelling and *in silico* trial simulations

4.2.5.1 General

Pharmacokinetic/pharmacodynamic (PK/PD) models[29][30] can usefully translate *in vitro*, non-clinical and clinical PK/PD data into meaningful information to support decision-making. At the individual level, substance PKs can either be described by non-compartmental analysis and compartmental PK modelling or by physiologically-based PK (PBPK) modelling. At the population level, population PK have become the most commonly used top-down models that derive a pharmaco-statistical model from observed systemic concentrations. PK/PD modelling involves on the one hand a quantification of drug absorption and disposition (PK) and on the other hand a description of the drug-induced effect (PD). PK/PD models and quantitative systems pharmacology (QSP) both aim for mechanistic and quantitative analyses of the interactions between a substance such as a drug and a specific biological system.[31]

PK and PBPK modelling are currently used for simulations for virtual patient populations in *in silico* clinical trials. The concept is that computer simulations are proposed as an alternative source of evidence to support drug development to reduce, refine, complement or replace the established data sources including *in vitro* experiments, *in vivo* animal studies and clinical trials in healthy volunteers and patients.

4.2.5.2 Challenges

The challenges are as follows:

— Reliable data sources for systems-related parameters are currently limited.

— Methods for data generation, collection and integration are not standardized.

— Reporting of results is very heterogeneous and inconsistent.[32]

— Tools to be used and criteria for model evaluation are very variable across projects.

— Very limited platforms (systems model) are currently considered reliable and qualified for regulatory submission.

4.2.6 Artificial intelligence models

4.2.6.1 General

Data-driven approaches, utilizing artificial intelligence (AI) and machine learning (ML) treat the mechanism as unknown and aim to model a function that operates on data input to predict the outcome, regardless of the unknown physiological processes. The mechanisms operating in the complex systems being modelled, i.e. which factors together drive outcomes, are considered too complex to be determined (black-box models). The quality of black-box models is assessed through the accuracy of their predictions, tested in a variety of ways. These data-driven models can be applied in a hypothesis-naive way, made as to which factors drive the causal mechanism.

ML approaches learn the theory automatically from the data through a process of inference, model fitting or learning from examples.[33] ML can be supervised, unsupervised or partially supervised (see Annex B).

4.2.6.2 Challenges

The challenges are as follows:

— Imprecise reporting, which makes it difficult to obtain the full benefit of results, navigate biomedical literature and generate clinically actionable findings.

— Data standardization, since most *in silico* methods require comparable input data.

— Data based on group associations, or pre-determined understanding of clinical relationships, can bias and limit AI/ML predictions (inappropriately pre-processed data).

— Different proprietary systems in healthcare information technology (IT) make data extraction, labelling, interpretation and standardization highly complex procedures (data lockdown).

4.3 Standardization needs for computational models

4.3.1 General

Major challenges in the field of personalized 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 personalized medicine requires a basic understanding of the modelled system, as well as of its biological and physiological background, and finally of the applied virtual experiments.

Because of the heterogeneous nature of the data in personalized medicine, harmonized strategies for data integration are required that utilize broadly applicable standards to allow for reproducible data exploitation to generate new knowledge for medical benefits. The two key components for which broad standardization efforts make most sense in the model building process are thus data integration and model validation (see Figure 2).

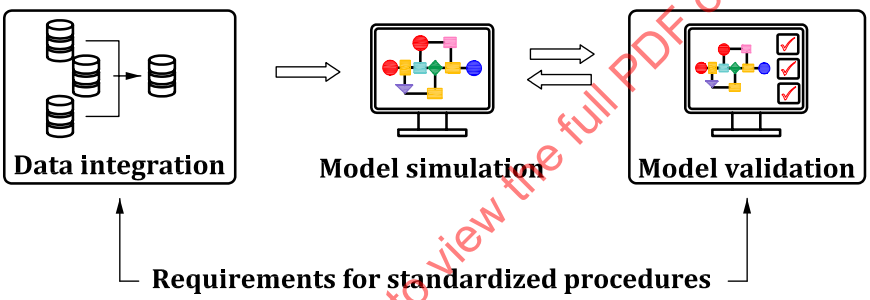

Figure 2 — Data integration and model validation as key factors for standardization requirements for computational models

4.3.2 Challenges

Although for many different data types used in personalized medicine there are domain-specific annotation standards and terminologies available (see Annex A), the process of model building possesses the following variety of challenges:

— High degree of variability regarding data types (structured versus unstructured, molecular, clinical, laboratory, patient-reported, etc.).

— Differences in coding and calculation within data types (between-machine variability, different measurements, etc.).

— Heterogeneous utilization of existing data.

— High effort of data harmonization in terms of time, resources and cost.

— Models relevant for clinical use need to be fit for purpose.

— Differences in IT systems used in data generation, e.g. enterprise resource planning systems and laboratory result software or hardware, at national, regional or clinical centre level.

— Adoption of different standards such as NPU (Nomenclature for Properties and Units) or LOINC (Logical Observation Identifier Names and Codes).

— Long-term variety and dynamics of data and standards.

— Differences in implementation of international terminologies such as the International Classification of Diseases (ICD).

— Language differences in unstructured text, and other factors.

4.3.3 Common standards relevant for personalized medicine

The use of common standards developed by specific user communities and different stakeholders, as well as standard-defining organizations, has been enhanced as they have been coupled to tools, which have spread in the respective field of research. Annex A provides an overview of some of these standards currently in use by different communities.

4.4 Data preparation for integration into computer models

4.4.1 General

Computational models in the life sciences in general and in personalized medicine research in particular are increasingly incorporating rich and varied data sets 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 personalized medicine, these data shall:

— include or be annotated with sampling and specimen data that follow the requirements and recommendations in accordance with the relevant domain-specific standards;

— be formatted using generally accepted and interoperable standard data formats commonly used for the corresponding data types (in accordance with 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 ISO 20691:2022, Annex B);

— follow best practice requirements and recommendations of generally accepted domain-specific data interoperability frameworks;

— be structured in a way that allows integration of the data into a model, 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.

4.4.2 Sampling data

Dedicated measures shall 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

cannot determine the situation in the patient, but determines an artificial profile generated during the pre-examination process.

NOTE Important measures include, for example, times and temperatures of sample transportation not exceeding the specifications provided in relevant International Standards (e.g. ISO 20916, ISO 20186-1) and International Technical Specifications (e.g. ISO/TS 20658), giving guidelines on all steps of the pre-examination workflow.

Conditions applied to a specimen shall be documented in addition to other important metadata, including but not limited to the content of Table 1.

Table 1 — Important metadata collected during pre-examination workflows

Metadata	Details
Specimen collection	— ID of responsible person
Information about specimen donor	— ID — Health status (e.g. healthy, disease type, concomitant disease, demographics such as age and gender) — Routine medical treatment and special treatment prior to specimen collection (e.g. anaesthetics, medications, surgical or diagnostic procedures, fasting status) — Appropriate consent from the specimen donor/patient
Information about the specimen, collection from the donor or patient and processing	— Type and the purpose of the examination requested — Specimen collection technique used (e.g. surgery, draw, flush) — Time and date when the specimen is removed from the body — Documentation of any additions or modifications to the specimen after removal from the body (e.g. addition of reagents)
Specimen storage and transport	— Temperatures of the collection device's surroundings
Specimen reception	— ID or name of the person receiving the specimen — Arrival date, time and conditions (e.g. labelling, transport conditions including temperature, leaking/breaking of the specimen collection container) — Nonconformities, including those from collection and transport requirements
Specimen processing and isolation of analyte	— Procedure and any modification applied from method referenced — Storage buffer of analyte
Information on isolated analyte	— Quantity and quality/integrity of analyte
Storage of isolated analyte	— Date and time of storage start — Temperature and method applied for storage

4.4.3 Data formatting

The first step in constructing a predictive computational model in personalized medicine is collating the data sets that need to be integrated into the model (which typically originate from various sources), and describing the different aspects of the studied subject and specimen to be modelled and simulated. The following different major resources for data in personalized medicine can be identified:

a) clinical data;

b) laboratory data (including omics data, as well as data from histology and cytology);

c) sample data and trial data;[34]

d) data from medical imaging, functional examinations and other clinical tests.

NOTE 1 Sensory data, either from medical or personal devices, are becoming more and more important and constitute another type of resource.

Each of these groups shall be structured in the corresponding formats in a way that allows conflation and, thus, ensures interoperability. The aim shall be to render the data interoperable, so that people (e.g. researchers) and also software tools can identify the key information in the data files/entries and interrelate corresponding pieces of information. To allow for seamless integration of data used for the construction or parameterization of such computational models, before their integration into the corresponding models, data sets shall be formatted using generally accepted, appropriate and interoperable standard data formats commonly used for the corresponding data types (in accordance with ISO 20691:2022, Annex A, with extensive examples of recommended standard formats).

The used data format and its version shall be unambiguously documented with the data to allow for later decoding of the data.

NOTE 2 It is often difficult for researchers to be able to identify suitable standards for use within their field, as availability is not enough: other researchers within the field also need to use the standards, and there has to be software to facilitate the generation and exchange of standardized data and models. Online resources are available for bundling information about available standards for formatting data in the domain. One widely used example for such a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies, is the publicly available FAIRsharing portal.[35][1)] Recommended formats referenced in ISO 20691 can be found online as an actively curated, constantly maintained and updated list as the ISO 20691 FAIRSharing Collection.[36]

4.4.4 Data description

NOTE Research data are distributed across resources and repositories hosted by different institutions, ranging from individual companies, institutions, research groups and universities, to national and international infrastructures with their data repositories. Many life science disciplines have developed their own, discipline-specific resources and catalogues with specific, sometimes implicit practices and standards. However, a standardized set of metadata has the potential to overcome these data silos.

Descriptive metadata are necessary to integrate biomedical, laboratory and clinical data into computer models. Metadata provides structured descriptions of the content, context and provenance of data sets. Descriptive metadata provides searchable information, making the data sets themselves discoverable and providing mechanisms for data citation. Metadata also enables users to judge whether a particular data set is suitable for their particular research purpose. Requirements for descriptive metadata can be found in domain-specific minimum information guidelines (see ISO 20691:2022, Annex B) or, more generally, in the FAIR principles for data stewardship.[7]

4.4.5 Data annotation (semantics)

For many data types used in personalized medicine, domain-specific annotation standards and terminologies are available. For example, UniProt[37] or the Protein Ontology[38] should 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[39] should be used to identify specific genes, or cellular components, whereas the Foundational Model of Anatomy (FMA)[40] should 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 are most commonly structured and codified by specific formatting standards for medical data. These can be the interoperability standard HL7 Fast Healthcare Interoperability Resources (FHIR),[41] or the standard for

1) This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product. Other web resources are also available for summarizing suitable formats and vocabularies.

electronic health records (openEHR[42]). Semantical content is usually annotated with domain-specific clinical terminologies, e.g. International Classification of Diseases (ICD)[43], Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) or LOINC (see Table 3). Thus, for semantic data annotation in personalized medicine there already are many formatting options, but existing annotation standards shall be consistently applied.

4.4.6 Data interoperability requirements across subdomains

4.4.6.1 General

A data interoperability framework should be defined and implemented to accommodate the heterogeneous and diverse nature of data in advancing science and knowledge where data, metadata and standards evolve rapidly, resulting in a dynamic complexity of data types and definition.

The life science domain is converging and over-reaching into other domains of science and engineering. Requirements such as those of international units of measurements or data schema specification are a shared commonality that should not be re-invented, but rather adopted to be practical and pragmatic. The metadata-crosswalk mechanism is a critical component to cross-subdomain data interoperability including data indexing facilitated by a schematic approach such as that of Schema.org[44] and Bioschemas[45].

Existing frameworks such as EOSC-Life Recommendations on FAIR metrics for EOSC,[46] and EOSC Enhance's analysis of research data cataloguing best practices[47] examine requirements of both generic and specific metadata standards for exchanging metadata information. Technical surveys published by collaborative initiatives such as these should be taken as a reference of metadata exchange best practice for reuse. Establishing a framework for data interoperability should clearly identify levels of integration, separating out interoperability at technical implementation level, organizational guideline level and research infrastructure policy level. Metadata properties should be examined carefully to identify appropriate data interoperability framework(s) to be (re)used. This is summarized in Table 2 by EOSC Enhance's analysis.[47]

Table 2 — Summary of metadata properties

Metadata property	Description
Descriptive	Information about the entities such as identifiers, type, location, etc., often including the use of controlled vocabularies for classification and indexing and links to related resources. A precise idea about the content of a resource so that users can find it and know if it is appropriate: a title, a description, keywords and one or many points of contact (creator, author, editor), etc. Also called "functional" metadata.
Technical	Technical processes used to produce or required to use a digital object: type of resource, encoding format, file size, software used. Also called "operational".
Administrative	Administration aspects such as intellectual property (IP) rights and acquisition, and information concerning the creation, alteration and version control of the metadata itself. Also called "operational".
Structural	Information about the files that make up the resource and the relationships between them.
Use	User access, user tracking and multi-versioning information.
Preservation	Documents actions such as migrations and checksum calculations. Also called "operational".

4.4.6.2 Challenges

The challenges are as follows:

— There is not a one-size-fits-all solution to identify a common interface of community standards due to the diversity in data types and data standards/metadata, and the multi-dimensional definition of what constitutes a data set.

— Data resources identified for interoperation and integration are developed at different maturity levels at any given point of time.

— Adoption of (meta)data standards for data interoperability is driven, not mutually exclusively, de facto and de jure. The practice is determined by scenario of a complex technical implementation environment.

— Data interoperability across subdomains can occur in all levels: top-level policy, intermediate-level implementation guidelines and recommendations, and operational-level technical development/ implementation.

— Mapping of metadata standards is critical to data interoperability, but a common model mapping of non-parallel orthogonal metadata standards is very difficult and expensive to achieve.

4.4.7 Data integration

Currently there are no widely accepted, overarching strategies or concepts to harmonize the integration of heterogeneous health and disease data for computational models for personalized medicine. However, the development of harmonized strategies for data integration is a key topic for standardization efforts in personalized medicine; standardization of input data is both feasible and necessary. Data harmonization is usually to some extent necessary for data integration and requires canonical interoperability (placement of the data in an ontology or structure), syntactic interoperability (data packaging) and physical interoperability. These requirements can be challenging, especially when legal restrictions apply or when handling very large data sets. When mapping data sets onto each other, sematic interoperability shall be a clear prerequisite for subsequent comparison and/or analysis.

Computational modelling of all kinds requires high quality data integration to create reliable output, but the vast quantities of research, medical and health data in existence are too disparate to be harnessed optimally. Data can be integrated at different levels, also with regard to personal identifiability. These represent the following different opportunities:

— Individual level integration: Data linked at the personal ID level can be used for personalized prediction of disease progression, i.e. personalized therapy based on past disease and health data. Individual level data integration is important for research in rare diseases, where patient numbers are low and counting a patient more than once affects data quality.

— Integration of variables: Data can be combined at the variable level, e.g. diagnoses, laboratory values, therapeutic intervention, symptoms, scores, outcomes or omics.

— Integration of unstructured data: Unstructured data can be integrated and processed together.

— Federation of data or validation of findings: Data sets can contribute to joint results by training an algorithm sequentially on the data sets without combining them, or by using new data sets for validation. While there are models that can process non-harmonized data and learn from them, this is often inefficient, and for most systems of federation of clinical data, e.g. Observational Medical Outcomes (OMOP) Partnership[48], data sets still have to be harmonized, and interoperable to return useful results.

4.4.8 Data provenance information

The complete history of the data should be documented using structured, interoperable and hence machine-actionable provenance information in accordance with ISO/TS 23494-1.

An uninterrupted chain of provenance information should be maintained by linking together metadata describing any preceding processing steps, methods, tools, biological entities, biological material and data utilized to generate the data documented.

NOTE A complete chain of provenance information enables both the assessment of data quality and its fitness for a particular purpose, and establishes reliability and reproducibility of the data by tracing its origin, generation, processing and analysis.

Appropriate precautions, such as access control mechanisms, shall be taken if the provenance information can contain sensitive or personally identifiable information.

A defined model, corresponding serializations and other supporting definitions to enable the interoperable interchange of provenance information in heterogeneous environments such as the web should be established and documented for all data in every format.

4.4.9 Data access

Harmonized strategies for data integration are key to standardization efforts in personalized medicine, which is critically dependent on efficient sharing of relevant data. However, in order to comply with relevant data protection and security, the majority of the data are classified as controlled access. As a consequence, over 1 000 different data access committees (DACs) and data access agreements (DAAs) from different jurisdictions are in use to control the access to data deposited in online resources, such as the European Genome-Phenome Archive (EGA).[49]

To address this limitation, harmonized data access agreements (hDAAs) shall be used to control data access by researchers and modellers for restricted data, such as the hDAA defined by EU-STANDS4PM for new submissions of controlled access data to EGA and similar archives (the current version of the EU-STANDS4PM hDAA is available for download from the EGA[50] and EU-STANDS4PM[51] websites).

NOTE 1 Such hDAAs stop the exponential proliferation of different access agreements and reduces the bureaucracy for both DACs and prospective data users, and eliminates the use of inappropriate clauses present in some DAAs. Hence, hDAAs for controlled access data facilitate improved data sharing and governance.

EXAMPLE The new EU-STANDS4PM hDAA is an agreement between the data provider (controller) and data user (recipient). It is executed via a DAC and defines the details of how the data can be used and stored by the user. The EU-STANDS4PM hDAA has been developed based on the European General Data Protection Regulation (GDPR) as interpreted by the EU-STANDS4PM framework of Ethical, Legal and Social Implications (ELSI), which is available as a full and compact version.[52]

NOTE 2 The EU-STANDS4PM hDAA is designed for use in conjunction with new controlled access data being deposited into EGA. It does not replace existing DAAs already linked to deposited data. Implementing the EU-STANDS4PM hDAA involves adding a link to one of the download options along with a link to the corresponding DAC to any new submission of controlled access data. A detailed workflow describing the entire process of submitting controlled access data in general is being developed by the International Human Epigenome Consortium (IHEC)[53] Bioethics Working Group[54] in collaboration with EU-STANDS4PM.[51]

4.5 Model formatting

If feasible (e.g. for mechanistic models), the encoding of a predictive computational model in personalized medicine shall be formatted using generally accepted, appropriate and interoperable standard model formats commonly used for the corresponding data types (in accordance with ISO 20691:2022, Clause A.3, with extensive examples of recommended standard formats).

The used data format and its version shall be unambiguously documented with the data to allow for later decoding of the data.

NOTE 1 The corresponding scientific communities have defined many domain-specific standards to consistently structure and format data, models and their metadata for modelling[55] and simulation[56] in the life sciences. These standardization efforts are driven by standardization initiatives, such as the Computational Modelling in Biology Network (COMBINE).[57][58][59] For providing the potential users with an overview and comparable information about such standards, web-based information resources have been developed and are publicly available, such as the NormSys registry[60] for modelling standards.

Examples of established model formats mainly used for models consisting of molecular entities and describing their interactions and dynamical interplay include the following:

— SBML (Systems Biology Markup Language) as standardized interchange format for computer models of biological processes[61];

— CellML, a standard format to store and exchange reusable, modular computer-based mathematical models;

— NeuroML (Neural Open Markup Language), which allows standardization of model descriptions in computational neuroscience;

— SBOL for synthetic biology models;

— BioPAX for biological/biochemical pathway models.[62]

NOTE 2 These are based on XML (extensible markup language)[63], a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. These formats represent the structure of a model (e.g. the biochemical network) and enable annotation of the model to better convey a model's intention.

4.6 Model validation

4.6.1 General

Any algorithm should perform 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.[64] There are guidelines and methods for validating accuracy and confidence in predictions of models.[65]

NOTE Common to all computational models is a need for validation[66] and accuracy, especially in the area of drug and medical device development.[67][65] However, in contrast to data input, model validation methods are considered to be individual and type-specific (see 4.6.2).

Performance evaluation of computational models should be transparent and consistent to existing guidelines. When alternative validation methods are chosen, the reasons should be explained.

4.6.2 Specific recommendations for model validation

It is important that recommendations for model validation (as well as for data integration) are commonly accepted by all the involved parties (clinicians, health technology assessment agencies, academia, industry, regulators and patient organizations, but also funding organizations such as the EU Commission).

Table 3 contains a summary of specific recommendations for different modelling approaches.

Table 3 — Specific recommendations for model validation

Model	Recommendations	
Cellular systems biology models	R1:	A standardized protocol for patient history and information should be developed and integrated with clinical standard protocols, e.g. electronic health records (EHRs) and FHIR.
	R2:	The use of patient expression profile, e.g. mRNA or protein expression, should be agreed upon.
	R3:	Prediction models with a fitting hypothesis should be developed.
	R4:	Model replication and reproduction should be used before considering clinical trials.
	R5:	Model predictions (e.g. molecular biomarkers selection) should be compared with established clinical ones.
	R6:	User-friendly graphical interfaces should be developed to ease the use of models in clinic.

Table 3 *(continued)*

Model	Recommendations	
Risk prediction for common diseases	R1:	The highest possible diversity of samples for all genetic studies on complex diseases, as well as for performing functional studies, should be ensured.
	R2:	The development of relevant transparent, standardized and detailed methods clearly stating methodological choices made with necessary justifications to enhance reproducible research should be enabled.
Disease course and therapy response prediction	R1:	Disease-specific scores including objective parameters of disease activity and progression should be harmonized.
	R2:	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 should be supported.
	R3:	Minimum clinical criteria for a systems medicine trial, taking harmonized scores and quantitative, validated patient-reported outcomes into account, should be defined.
	R4:	Concepts for integrating model-based prediction and AI content into curricular education in medicine and medical life sciences should be developed.
	R5:	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 (the current schemes of the Innovative Medicine Initiative can serve as a starting point), should be commenced.
	R6:	New models of public-private partnerships for a strong European healthcare economy on IP participation and protection of mutual interests should be developed.
Pharmacokinetic/ -dynamic modelling and *in silico* trial simulations	R1:	Appropriate *in vitro-in vivo* extrapolations from targeted assays to aid in validating PK/PD models (a full validation of PK/PD models requires comprehensive measurements of the drug PK, as well as of the resulting therapeutic effect) should be developed.
	R2:	An agreement among relevant stakeholders on required criteria for model adequacy for a well-defined context of use should be implemented.
	R3:	It should be ensured that the standards (to be) used for model development, evaluation, reporting and related decision-making are commonly acknowledged by all the involved parties (regulators, health technology assessment (HTA) agencies, academia, research centres, industry and patient organizations) are that they are relevant for all the types of models that can be used. NOTE Of interest is the standard recently published by the American Society of Mechanical Engineers (ASME) VV-40-2018 "Verification and Validation in Computational Modeling of Medical Devices" which was proposed to be extended to other biomedical products such as drug and combined medical products[68] and to Reference [69]. In the current situation, a similar initiative oriented to disease description and drug development would be of great value.
	R4:	The requirements for adequate implementation of the two other components (patient's and other design characteristics) should also be standardized and discussed by the involved stakeholders.
Artificial intelligence models	R1:	Model validation should involve the three-phase process, as detailed in following recommendations R2 to R4.
	R2:	In the first phase, outcome prediction for the AI model should be compared to standard measures that employ best clinical practice and using established evidence-based hierarchy which takes into account historical practice, clinical trials and systematic review/meta-analyses.[70]

Table 3 *(continued)*

Model	Recommendations	
	R3:	If the AI model shows a predictive advantage, a second phase of validation should involve use of a High Confidence Off Policy Evaluation method as described previously[71] where the AI model is compared to clinician decision-making.
	R4:	If beneficial, phase 3 should provide clinicians access to AI prediction points.

4.7 Model simulation

4.7.1 General

Model simulation brings the models to life. Models used in personalized medicine vary, based on the large spectrum of mathematical concepts (e.g. graph theory, dynamical systems theory). Consequently, model simulation, computational costs and the resolution of results is model-specific with models having different levels of abstraction, and predictive power that can be categorized into three main levels: topological, constraint-based and kinetic modelling. A description for the three main methods and their simulation comparison are listed in Table 4.

NOTE 1 Topological analysis is a graph-based approach that has been introduced for simplification, reconstruction, analysis and comprehension of complex behaviour in biological systems. Broadly, topological approaches can be used to explore:

— collective behaviours (e.g. scale-free properties of a network);

— sub-network behaviours (e.g. functional motif discovery);

— individual behaviours (e.g. prioritization of influential nodes) of various networks.

NOTE 2 A constraint-based modelling approach calculates the flow of metabolites through a metabolic reaction. This method allows prediction of disease mechanism and drug target prediction. Likewise, the topological analysis applies to large-scale systems with low computational costs, but the results are quantitative.

NOTE 3 Kinetic modelling is an ordinary differential equation-based approach, which analyses the behaviour of a biochemical reaction over time. This approach usually applies for small-scale systems (e.g. a single pathway). Although this method is fully parameterized, it needs many experimental data including enzyme kinetic law and sample concentration.

Table 4 — Important methods for model simulation

Method	Simulation	Output	Advantages	Limitations
Topological modelling	Graph-based analysis	— Centrality indices — Motifs — Clusters	— Analysing large networks — Low computational demand	— No compartment — Qualitative results — Examines the static aspects of the model
Constraint-based modelling	Flux balance analysis	— Essential genes — Robustness analysis	— Analysing large networks — Quantitative results — Compartmental — Low computational demand	— Steady-state approximation — No diffusion
Kinetic modelling	Deterministic or stochastic approaches	— Time-course analysis	— Considering the dynamics of a system — Quantitative results — Fully parameterized	— Analysing small networks — High computational demand — Requires a large amount of experimental data

4.7.2 Requirements for capturing and sharing simulation set-ups

Models can be used for more than one investigation, and they can be linked to multiple data sets. Sometimes, a model undergoes complex pre-processing before the actual simulation is executed. Therefore, simulation set-ups and possible multiple initial model conditions shall be captured. These pieces of information should be kept separately from the actual model code, making the model code a template for studying a biomedical system. The simulation set-ups then specify how this template can be used to simulate varying conditions and hypotheses.[72]

A recognized standard shall be used for capturing and reporting simulation set-ups and corresponding information (in accordance with ISO 20691:2022, Clause A.5).

The minimum information about a simulation experiment (MIASE) recommendations[73] describe the minimal set of information that should be provided to make the description of a model simulation available to others. This information should include the list of models to use and their modifications, all the simulation procedures to apply and in which order, the processing of the raw numerical results, and the description of the final output.

NOTE 1 MIASE allows for the reproduction of any domain-specific simulation experiment. The COMBINE community recommends the use of the Simulation Experiment Description Markup Language (SED-ML) for encoding of simulation set-ups,[56] and the Kinetic Simulation Algorithm Ontology and the Systems Biology Ontology[74] for encoding information on applicable simulation algorithms and mathematics. SED-ML encodes in a computer-readable exchange format the information required by MIASE to enable reproduction of simulation experiments. It has been developed as a community project. It is defined in a detailed technical specification and additionally provides an XML schema.

NOTE 2 The COMBINE community has developed the COMBINE archive format[75] that facilitates the reproduction of modelling and simulation experiments in biology by embedding all the relevant information in one file. A COMBINE archive is a single file bundling the various documents necessary to reproduce a simulation study, providing a model and all associated data and procedures. This includes for instance, but is not limited to, simulation experiment descriptions, all models needed to run the simulations and associated data files. The archive is encoded using the open modelling exchange format (OMEX).

4.7.3 Requirements for capturing and sharing simulation results

A recognized standard format shall be used for reporting and sharing the output results of a model simulation (e.g. in accordance with ISO 20691:2022, Clause A.5). The results shall clearly define the output of the simulation, in the sense that it specifies what shall be recorded and/or plotted in the output.

The simulation results shall be reported together with information on the simulation and analysis workflow that produced the results, as well as instructions how the results should be presented. Simulation results should be linked to information on the simulation set-ups that led to the corresponding output.

The raw simulation result sometimes does not correspond to the desired output of the simulation, e.g. it can be necessary to normalize a plot before output, or apply post-processing to the raw data. Any post-processing that needs to be applied to the simulation result before the final simulation output shall be documented together with the simulation output.

NOTE The simulation experiment description markup language (SED-ML)[56] encodes in a computer-readable exchange format the information required by MIASE (see 4.3.1) to enable reproduction of simulation experiments.

4.8 Requirements for model storing and sharing

Predictive computational models for research in personalized medicine shall be stored and shared in a way that makes them accessible to both humans with vested interest and machines via automatized workflows.

NOTE 1 Access to sensitive and/or confidential data can be restricted to comply with national law.

For potential re-use of the models, they shall be encoded in generally accepted, appropriate and interoperable standard model formats commonly used for the corresponding model type (see 4.5), providing information about model provenance, versioning and licences that apply, as well as about ownership and attribution to the model creator(s).

To provide the necessary metadata describing model assumptions, coverage and history, as well as the semantics and context of the model and its contained components, entities, relations and processes, these metadata shall be stored and shared with the respective models. Models of biochemical systems in personalized medicine shall be documented and annotated, e.g. according to the MIRIAM guidelines[76] as recommended by COMBINE[58][59].

It is recommended to store and share the encoded model together with or linked to its description (e.g. the publication text if shared via a published manuscript).

NOTE 2 Several model repositories, such as BioModels[77], JWS Online[78] or the Physiome Model Repository[79], offer model management tasks from model curation to model verification, model publication and model version control.

4.9 Application of models in clinical trials and research

4.9.1 General

Despite the immense potential offered by computational models and big data analytics for personalized medicine, until today, researchers and clinicians have dealt with significant barriers to the successful translation of this technology from bench to bedside. More precisely, the characterization of unknown model parameters, the need for models to faithfully reflect *in vivo* conditions, and the limited availability of fitting validation data to check models' accuracy and assess their reliability pose major obstacles in the path towards their clinical translation.

A standard plan to develop computational models for predicting biological outcomes shall be followed, i.e. the adoption of a precise stepwise strategy focused on three key points:

— choosing the data set;

— selecting the algorithm and training it to develop a prediction model;

— testing it in unseen data sets.

4.9.2 Specific recommendations

Table 5 contains a summary of specific recommendations.

Table 5 — Specific recommendations for the application of models in clinical trials and research

Recommendations	
R1:	Specific areas where models can be applied in clinical trials *a priori* in a stakeholder-based process should be defined, including patients/patient organizations to define in which area they would feel safe that such models are applied.
R2:	Discussion forums for stakeholder interactions on clinical outcome parameters and application of models should be established.
R3:	The development and definition of specific areas should be defined where models are applied to create blueprint solutions for individual trial schemes.
R4:	A framework endorsed by all the relevant stakeholders to support the model evaluation rationale should be defined.
R5:	Independent of the model type, the rigor in model evaluation should be the same, whereas all stakeholders (e.g. regulators, clinical investigators, scientists) should use the same validation tools.
R6:	Regulatory criteria that models have to fulfil to be applied in clinical trials should be defined.
R7:	Reporting schemes for model parameters and results of clinical trials should be defined.

4.10 Ethical requirements for modelling in personalized medicine

Both the development and the application of computational models give rise to ethical issues and shall take into account ethical standards regarding research ethics, patients' rights, data protection and privacy rights. In addition, risks of discrimination, as well as issues related to transparency, trust and accountability shall be equally addressed[80].

While developing computational models, adherence to ethical principles and requirements, such as respect for persons (autonomy), balancing risks and benefits, fair distribution of risks and benefits (justice) should be ensured. These principles are particularly relevant when vulnerable populations are involved. In addition, such activities should be performed by qualified staff and approved by ethics committees, as well as relevant competent and appropriately qualified physicians or other healthcare professionals.[81][82]

NOTE For general guidance on corporate responsibility principles, see ISO 26000.

While it is still under discussion how and to what extent consent should be applied as a legitimate basis for processing personal data, it should be noted that informed consent in terms of research ethics is still one of the fundamental research ethics requirements with only limited exceptions.

Transparency of models, in the form of interpretability or post-hoc explanations, is important for validation. Where the model is a complete "black box", and its logic entirely inscrutable, a possible recommendation is that it can be more ethically suitable for use in research and development than actual deployment in medicine.

Annex A
(informative)

Common standards relevant for personalized medicine and *in silico* approaches

Table A.1 — DNA, RNA, protein sequence formats

Standard	Description
FASTA	Widely used for representing nucleotide sequences or amino acids, developed for use in the FASTA program.[83][84] The FASTA format is simple and lacks facility for extensive annotation.
Sequence Alignment/Map (SAM) and Binary Analysis Map (BAM) format	Capture of sequences 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.[85] 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.[86]
ISO/IEC 23092 series (MPEG-G)	The ISO/IEC 23092 series (MPEG-G) 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 six parts: ISO/IEC 23092-1, ISO/IEC 23092-2, ISO/IEC 23092-3, ISO/IEC 23092-4, ISO/IEC 23092-5 and ISO/IEC 23092-6[2)].
General feature format (GFF)	Stores DNA, RNA or protein genetic sequence data.[87] 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.[88] Originally designed 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. Also designed to represent incomplete/partial sequences and relative ordering of parts in a genetic design.

2) Under preparation. Stage at the time of publication: ISO/IEC FDIS 23092-6:2023.

Table A.2 — Mass spectrometry

Standard	Description
mzML	Stores the spectra and chromatograms from mass spectrometry in an XML format. Now a well-tested open-source format for mass spectrometer output files that is widely used.[89]
mzTab	A more easily accessible format, for use with R or Microsoft Excel[a] tools in the field of proteomics and metabolomics. mzTab files can contain protein, peptide and small molecule identifications. In addition, they can contain experimental metadata and basic quantitative information.[90]

[a] Excel is the trademark of a product supplied by Microsoft. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named. Equivalent products may be used if they can be shown to lead to the same results.

Table A.3 — Medical imaging, digital imaging and communications in medicine

Standard	Description
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). A living standard that is regularly maintained and modified.[91] Adopted as ISO 12052.
European Data Format (EDF)	A standard to archive, share and analyse data from medical time series.[92]

Table A.4 — Semantic integrations

Standard	Description
BRIDG (Biomedical Research Integrated Domain Group Model)[93]	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, 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 artefacts. 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. The BRIDG Model is specified in ISO 14199.
HL7 FHIR (Fast Healthcare Interoperability Resources)	Designed to enable information exchange to support the provision of healthcare in a wide variety of settings. The specification builds on and adapts modern, widely used RESTful practices to enable the provision of integrated healthcare across a wide range of teams and organizations. The intended scope of FHIR is broad, covering human and veterinary, clinical care, public health, clinical trials, administration and financial aspects. The standard has been developed by an international group of people and is intended for global use and in a wide variety of architectures and scenarios.

Table A.4 *(continued)*

Standard	Description
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.[94] The primary labels use medical terminology used by clinicians and researchers. These are complemented by 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.[95]
Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[96]	SNOMED is a family of medical terminology systems. Originally conceived as a nomenclature, the latest version of the SNOMED CT can best be characterized 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.
Logical Observation Identifier Names and Codes (LOINC)[97]	Common language (set of identifiers, names, and codes) for clinical and laboratory observations. LOINC is a catalogue of measurements, including laboratory tests, clinical measures such as vital signs and anthropomorphic measures, standardized survey instruments, etc. LOINC enables the exchange and aggregation of clinical results for care delivery, outcomes management and research by providing a set of universal codes and structured names to unambiguously identify things that can be measured or observed. LOINC is recommended (also by HL7 and DICOM) for the exchange of structured documents (CDA) and messages.
ISO/IEEE 11073 series	Personal health data (PHD) standards: A series of standards addressing the interoperability of personal health devices (PHDs) such as weighing scales, blood pressure monitors, blood glucose monitors, etc.

Annex B
(informative)

Information on modelling approaches relevant for personalized medicine

B.1 Risk prediction for common diseases

Common complex diseases are multifactorial and polygenic; there are multiple genetic and environmental (e.g. diet, smoking) factors that affect an individual's risk of having a disease. The polygenic model assumes 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. However, to improve detection of genetic associations with diseases of interest, additional factors such as environmental exposure, epigenetics or parent-of-origin effects need to be considered in genetic studies. Taking these factors into account when determining the genetic risk factors for common diseases is critical for the estimation of polygenic risk scores.

Prediction based on AI (see 4.2.6) analysis of phenotypic information can be increasingly accurate.[98] However, there is also a more narrow definition of AI risk prediction, in which the expression prediction is reserved for cases of direct causality, and focuses on polygenic risk scores (PRS)[99][100][101].

B.2 Artificial intelligence models

Artificial intelligence models can be grouped in three categories: supervised, unsupervised and semi-supervised learning models.

Unsupervised learning comprises dimensionality reduction, permitting feature elicitation, compression and big data visualization, all of which can allow for better understanding of big medical data and the factors driving disease initiation and progression. For general understanding of the relationships between disease and co-factor, unsupervised learning offers the ability to discover new knowledge. More importantly, for the individual patient, unsupervised learning can uncover previously unrecognized phenotypes, which then can be used to refine outcome prediction. Unsupervised learning has the potential to take all features into account to identify specific clusters for precision as well as personalized medicine.

In supervised learning, the model is supplied with (labelled) input features that are considered when predicting a predetermined outcome from new data, either through regression (for continuous results such as number of days or months before disease debut) or classification (for discrete results such survival/death, or for image classification).

Semi-supervised learning employs both labelled and unlabelled data for training. When incorporating full data sets of disease comorbidities, supervised AI clinical decision support tools can be used to tailor predictions to the individual patient, according to labelled or unlabelled input data. However, it is important to always be aware of potential errors and biases being introduced by input data, when it is difficult to examine the mechanisms of the model.

B.3 Minimum amount of information to understand a model

There are two relevant checklists that serve as guidelines on the minimum amount of information required to understand a model. The minimum information requested in the annotation of biochemical models (MIRIAM)[76] focuses on information about data and models, required author information and necessary metadata in the form of semantic annotations. These annotations can improve the