TOWARD GOOD SIMULATION PRACTICE: BEST PRACTICES FOR THE USE OF COMPUTATIONAL MODELLING & SIMULATION IN THE REGULATORY PROCESS OF BIOMEDICAL PRODUCTS

Version: R6 – prefinal, open to public comments

Date: 06-05-2023

Table of Contents

1. GLOSSARY

Abbreviations, Acronyms and Specific Terms	Definitions
AFAP	As Far As Possible
ALARP	As Low As Reasonably Practicable
API	Application Programming Interface
ASME	American Society of Mechanical Engineers
ATMP	Advanced Therapeutic Medicinal Products
CAPA	Corrective Action and Preventive Action
CM&S	Computer Modelling and Simulation
CoU	Context of Use
CRO	Contract Research Organisation
DMO	Digital Mobility Outcome
DPIA	Data Protection Impact Assessment
EAA	Early Awareness and Alert
EMA	European Medicines Agency
FDA	United States of America Food and Drug Administration
FFR	Fractional Flow Reserve
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
GLP	Good Laboratory Practice
GMP	Good Manufacturing Practice
GSP	Good Simulation Practice
HIPAA	Health Insurance Portability and Accountability Act
HTA	Health Technology Assessment
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IMDRF	International Medical Device Regulators Forum

ISO	International Organization for Standardization
ISW_CoP	In Silico World Community of Practice
MDR	Medical Device Regulation [Regulation (EU) 2017/745]
MDRO	Multi-Drug Resistant Organism
MID	Minimal Important Difference
MMS	Method of Manufactured Solutions
NASA	National Aeronautics and Space Administration
NGS	Next Generation Sequencing
OECD	Organisation for Economic Co-operation and Development
PCP	Pre-Commercial Procurement
PMA	Pre-market approval
PPI	Public Procurement of Innovative solutions
QoI	Quantity of Interest
QSAR	Quantitative Structure-Activity Relationship
RCT	Randomised Controlled Trial
RTM	Requirements Traceability Matrix
RTM	Requirements Traceability Matrix
SaMD	Software as a Medical Device
SDP	Software Development Plan
SDP	Software Development Plan
SLC	Software Life Cycle
SOP	Standard Operating Procedure
SQA	Software Quality Assurance
SRD	System Requirements Document
SSED	Summary of Safety and Effectiveness Data
UML	Unified Modelling Language
VPH	Virtual Physiological Human
VV-40:2018	ASME standard “Assessing Credibility of Computational Modeling and Simulation Results through Verification and Validation: Application to Medical Devices”
VVUQ	Verification, Validation and Uncertainty Quantification

2. INTRODUCTION

Chapter Editors:

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Liesbet Geris (VPH Institute)
- Luca Emili (In Silico Trials Technologies)

Contributors (random order):

- Axel Loewe (Karlsruhe Institute of Technology)
- Bernard Staumont (University of Liège)
- Raphaëlle Lesage (VPH institute)
- March Horner (Ansys Corp.)
- Enrique Morales Orcajo, (Ambu A/S)
- Martha De Cunha Maluf-Burgman (Edwards Lifesciences)

2.1. Scope of this document

The term GxP indicates a collection of good practices, e.g., quality guidelines, to ensure a product is safe and meets its intended use. The most important examples of GxP in biomedicine are Good Laboratory Practice (GLP) and Good Clinical Practice (GCP). The GLP were developed by the Organisation for Economic Co-operation and Development (OECD); they provide "a managerial quality control system covering the organisational process and the conditions under which non-clinical health and environmental studies are planned, performed, monitored, recorded, reported, and retained (or archived)"[1]. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) curates the GCP. GCP provides an international ethical and scientific quality standard for clinical trials to facilitate the regulatory authorities' mutual acceptance of clinical evidence in the various ICH regions. GxP guidelines are available for various industries, including foods, medical products, medical devices, and cosmetics. In some cases, the GxP simply expresses best practices within an industrial sector; in others, they are elevated to quasi-regulatory standards, which must be met to achieve specific regulatory approval.

The use of Computer Modelling and Simulation (CM&S) in clinical medicine is usually referred to as *In Silico Medicine*. The term was first used in PubMed in 2013 and has become popular since then. The academic research community loosely uses the term *in silico methodologies* to indicate the use of CM&S to assess the safety and/or efficacy of new healthcare products, whether medical devices, medicinal products, or others. The term appeared in PubMed in 2002 (Ashelford et al., 2002). One of the issues with this term is that it uses the term "trial" loosely, whereas in the regulatory domain, the term is used in a much more specific way. To avoid such confusion, going forward, we will use the term *in silico* methodologies only in a colloquial way. Instead, we will use the term ***In Silico Methodology*** to indicate any use of CM&S as, at any level, a regulatory decision support tool on new medical products for which a marketing authorisation is requested, whether medical devices, medicinal products, or others.

[1] https://www.oecd.org/chemicalsafety/testing/overview-of-good-laboratory-practice.htm

This position paper on Good Simulation Practice (GSP) does not emerge from a vacuum. For example, since 2002, at least 21% of 565 original premarket approval (PMA) applications for medical devices had computational modelling efforts provided in the Summary of Safety and Effectiveness Data (SSED) (Morrison et al., 2019). Our community of practice, in general, and major regulatory agencies, in particular, have been reflecting on using predictive models as a development and de-risking tool for medical products. In some cases, such reflections took the form of guidance documents or technical standards for specific uses. In Annexe 1, we review the existing regulatory guidance on the topic.

While the regulatory community is actively engaged in developing a comprehensive regulatory framework that includes the use of computer simulations to support a medical decision with the introduction of the concept of "Software as a Medical Device" (SaMD), a similar level of engagement has been so far absent for the broader application of CM&S in regulatory decision-making processes. There is only one detailed resource for validating in silico methodologies applied to medical devices: the American Society of Mechanical Engineers (ASME) Verification & Validation (V&V)-40 standard[2], originally published in 2018, whose original scope was limited to medical devices (hereinafter referred to as VV-40:2018).

While the VV-40:2018 standard is a valuable resource, the authors of the present document believe there is a need for a document that summarises the good practices in using *in silico* methodologies in the regulatory process for all kinds of medical products. Such a document could play a role similar to that of the Good Clinical Practice (GCP), the Good Laboratory Practice (GLP), or the Good Manufacturing Practice (GMP) guidelines. Thus, by analogy, it could be named "Good Modelling & Simulation Practice for medical products", and hopefully, it may be curated and/or adopted by the members of the International Medical Device Regulators Forum (IMDRF). A GxP may remain a voluntary guideline or be elevated to a standard by standardisation bodies such as the International Council for Harmonisation (ICH) or the International Organization for Standardization (ISO). However, the compilation of good modelling & simulation practice for medical products is now a challenging task. *In silico* methodologies have started to be adopted only recently, and the experience is limited. Also, the expertise required to write such document is extremely multidisciplinary.

The VPH Institute[3] and the Avicenna Alliance[4] are two international not-for-profit organisations that represent all practitioners in the field of *in silico* medicine: the first represents the academic community, the second the industrial community. The EU-funded In Silico World project[5] operates, an online forum, in collaboration with the VPH Institute and the Avicenna Alliance, called In Silico World Community of Practice (ISW_CoP)[6]. The over 500 experts that participate in this ISW_CoP share a common professional or educational interest for *in silico* medicine. Within this community a consensus emerged on the opportunity to collaboratively compile a position paper aimed to summarise the current thinking within the ISW_CoP on the good practices for *In silico* methodologies, so as to provide a basis for the future development of a formal standard on the good modelling & simulation practice for medical products.

[2] https://www.asme.org/codes-standards/find-codes-standards/v-v-40-assessing-credibility-computational-modeling-verification-validation-application-medical-devices

[3] https://www.vph-institute.org/

[4] https://avicenna-alliance.com/

[5] https://insilico.world/

[6] https://insilico.world/community/

Thus, the scope of this document is to provide a list of the best practices on the use of computer simulation in assessing the safety and efficacy of medical products, as emerged by a consensus process within our ISW_CoP. The form we chose is a "Position Paper" - a public document providing an expert opinion to orient policies or standards. In this sense, the present document is not binding and represents only the consensus among some field experts. However, we hope this document might provide a starting point for a future standardisation effort by an appropriate body.

CM&S is used over the entire life cycle of medical products, including discovery and design, verification, development, optimisation, re-design, etc. However, this position paper focuses only on their use to assess the safety and efficacy of medical products.

The first output that the ISW_CoP produced was a systematic analysis of all possible Contexts of Use (CoU) for In Silico methodologies (Viceconti et al., 2021a). CoUs are concise descriptions of how the new methodology is intended to be used in medical products' development and regulatory assessment process.

To organise this long list, we used the taxonomy presented in Table 1.1, which will also be used throughout this position paper. The safety and efficacy of medical products are usually investigated using experimental methodologies: ***in vitro* and *ex vivo*** experiments, ***in vivo* animal** experimentation, or ***in vivo* human** experimentation. *In silico* methodologies are a valid alternative to these experimental methodologies. Using terminology that was first used to categorise alternatives to animal experimentation, In Silico methodologies can be used to **reduce** the experiment (fewer bench tests, fewer animals enrolled, fewer patients enrolled), **refine** the experiment (reduce the suffering of animals, reduce risks for humans, improve the ability of pre-clinical studies to predict the clinical outcome, generalise the experimental finding, etc.), and **replace** the experiment (replace the experiment entirely). This produces a 3x3 taxonomy (table 1.1), which will be used in the remainder of this document.

Table 1.1. Taxonomy of *in silico* methodologies

	Reduce	**Refine**	**Replace**
Preclinical In Vitro/Ex Vivo Experiments	Reduce the number or duration of in vitro/ex vivo experiments	Improve the predictive accuracy of safety and/or effectiveness provided by the in vitro or ex vivo experiment	Replace a portion or all the required in vitro or ex vivo experiments
Preclinical Animal Experiments	Reduce the number of animals involved in the experiment, or its duration (adoption of sustainability principles)	Alleviate the suffering of the animals involved, or improve the predictive accuracy of the safety and/or effectiveness provided by the animal experiment (solving or acknowledging animal protection issues)	Replace animal experiments used for the prediction of the expected safety and/or effectiveness of a new treatment during clinical experimentation
Clinical Human Experiments	Support the design of clinical experiments. Reduce the number of clinical studies, their duration, or the number of subjects involved. Solving scarcity on patients population related to rare diseases and where patients are children.	Reduce the risks for the humans involved or improve the predictive accuracy of the safety and/or effectiveness provided by the human trials.	Replace human experiments used for the prediction of the expected safety and/or effectiveness of a new treatment.

2.2. The critical elements of a Good Simulation Practice standard

Chapter 3 treats the topics considered essential in future GSP standards separately. Here we provide a brief introduction and summary.

2.2.1. Theoretical foundations of Good Simulation Practice

Regulatory science focuses on problems very close to clinical application. Thus, in general, its practitioners are not interested in the more fundamental aspects treated by mathematics, philosophy of science, and epistemology (study of human knowledge). However, the extreme interdisciplinarity involved with computer modelling and simulation in the development and de-risking of medical products makes it difficult for every single group of experts to use the epistemological guidelines accepted and established in the practice of their discipline. Having solid theoretical foundations helps in these cases to find common ground across different disciplines and epistemologies. The goal of Chapter 3 is to provide such foundations.

2.2.2. Model development

A computer model is, first and foremost, a software artefact; as such, it must be developed and tested using the quality assurance principles in software engineering. While this is a relatively mature topic for regulatory science, which has been specialised for biomedical applications with the introduction of the so-called software as a medical device category of medical devices, there are some specificities of providing quality assurance for software with predictive purposes that require specific treatments in a future GSP standard. In Chapter 4, we analyse this topic in full detail.

2.2.3. Model credibility

Even if a model has been developed with the highest possible quality standard, this does not guarantee that the predictions this model provides can be trusted per se. The problem of assessing the credibility of a model's prediction is a problem that has been addressed in the regulatory science of high-risk products such as nuclear power plants or passenger aircrafts. Yet, in the biomedical domain, this is a very recent topic.

Annexe 1 provides an overview of all regulatory documents that address this problem. Still, even the most recent efforts, such as the ASME VV-40:2018, leave an ample portion of the territory untouched. VV-40:2018 targets the development of medical devices, leaving out drug development and the development of ATPs. The classic VVUQ approach the VV-40:2018 refers to is robustly defined for purely mechanistic models, models built exclusively from widely accepted theories; however, many predictors are now built using data-driven methodologies, where no theory is involved. Furthermore, in practice, most models are called *grey-box models* because they are built by combining mechanistic and empirical knowledge. In Chapter 5, we provide a systematic discussion of the topic.

2.2.4. Possible regulatory pathways

The regulatory assessment of *In silico* methodologies does not fit well with the traditional separation between drugs and medical devices. It must include elements of technical validation more common in the regulatory pathways of medical devices, but also elements of clinical validation more common in the regulatory pathways of medicinal products. In Chapter 6, we explore the issue of which regulatory pathway is most suitable to qualify in silico methodologies to be used in the regulatory

assessment of new medical products. We describe four possible pathways and discuss their pros and cons.

2.2.5. Possible Health Technology Assessment pathways

In silico methodologies can play an essential role in the marketing authorisation of new medical products, their cost-benefit assessment, the definition of prescriptive appropriateness, and post-marketing surveillance. In Chapter 7, all these aspects are considered and discussed with concrete examples.

2.2.6. Ethical review of in silico methodologies

Before it starts, any experimental study on humans must be reviewed by an independent organisation known in Europe as Independent Ethics Committee and in the USA as Institutional Review Board. Chapter 8 explores if and how such a review process needs to change when *in silico* methodologies are involved.

2.2.7. The role of the sponsor in in silico methodologies

The sponsor is "an individual, company, institution, or organisation which takes responsibility for initiating, managing, and/or financing a clinical trial"[7]. The sponsor plays a vital role in conventional trials, codified in detail in various standards and guidelines, such as the Good Clinical Practice. Chapter 9 explores how such a role needs to be extended when *in silico* methodologies are involved.

2.2.8. The role of the Investigator in in silico methodologies

Another role that needs to be partially redefined when the clinical evaluation of a new medical product involves *in silico* methodologies is that of the Investigator. In a clinical study, the Investigator is the person involved in running the study. The Investigator may help prepare and carry out the protocol (plan) for the study, monitor the safety of the study, collect and analyse the data, and report the results of the study. When *in silico* methodologies are involved, the Investigator is also responsible for carrying out the modelling tasks and generating the *in silico* evidence. Chapter 10 explores how these additional responsibilities change the Investigator's profile and role.

2.3. Essential Good Simulation Practice recommendations

- In Silico methodologies can be categorised depending on how they are used as an alternative to experimental methodologies: to refine, reduce, and replace *in vitro*, animal, or human experimentation.

[7] https://toolkit.ncats.nih.gov/glossary/clinical-study-sponsor/

3. THEORETICAL FOUNDATIONS OF GOOD SIMULATION PRACTICE

Chapter editors:

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Miguel Juarez (University of Sheffield)

Contributors (random order):

- Marc Horner (Ansys)
- Daniela Calvetti (Case Western Reserve University)
- Erkki Somersalo (Case Western Reserve University)
- Liesbet Geris (University of Liege)
- Axel Loewe (Karlsruhe Institute of Technology)
- Martha De Cunha Maluf-Burgman (Edwards Lifesciences)

3.1. Introduction

This position paper aims to support future standardisation efforts on Good Simulation Practice. Good practice standards are usually the summary of best practices, collected empirically and consolidated through consensus among practitioners. As such, they are the least theoretical artefact one can expect in regulatory science. Thus, it might require some explanation on why we decided to add a chapter on some of the theoretical foundations supporting the concepts in the following chapters.

Regulatory best practices emerge through consensus among practitioners. This implies that such practitioners are culturally relatively homogenous and share the same vocabulary. Even more important, they share a common *epistemology*, the principle around which humans establish new knowledge, in this case, knowledge on the safety and efficacy of new medical products. This is one of the reasons why the regulatory assessment of medicinal products and medical devices remain separated, despite the more frequent combination products; each class of products has its own vocabulary, expertise, and epistemology.

Nevertheless, there are also commonalities. For example, the whole regulatory science is formulated as purely empirical, where experimental evidence and even better real-world observations are considered the only source of reliable information. Introducing modelling and simulation in the regulatory process raises several epistemological challenges. Evidence is predicted, not observed. Such predictions can be based on well-accepted theories that resisted extensive falsifiability efforts, theories that might still be debated, and even purely phenomenological observations on a large volume of observational data. It is quite clear that a predictive model and a controlled experiment are different ways to investigate physical reality, but how they differ is debatable. Even more complex is the definition of a formal process to establish the truth content of a model's prediction (what we call here "credibility").

Last but not least, the introduction of computer modelling and simulation must add to the panels of experts that develop by consensus the good simulation practice totally new expertise such as applied mathematics, computer science, software engineering, and a whole territory of engineering science sometimes referred to as Modelling and Simulation in Engineering. But this creates a group of experts

with different backgrounds, terminologies, and even epistemologies. This is why the discussion around the regulatory acceptance of *in silico* methodologies is so complex; the involved experts struggle to communicate and collaborate effectively.

There is no easy solution to this problem. People with different expertise and backgrounds will have to try to talk to each other and try to understand the other points of view. But in such a complex debate, we believe it is essential to have some theoretical foundations to which we can resort when the discussion reaches a dead end.

Thus, contrary to all others, this chapter does not directly contribute to the regulatory science debate on the GSP. As such, it might not be of particular utility to the regulators, although it may serve as an indirect nexus between the regulatory and the CM&S sciences. However, we believe it is a necessary element of such a document, and it might come in handy in some complex discussions that the consensus process will inevitably impose.

3.2. What is a model in science?

"A model is an invention, not a discovery" (Massoud et al., 1998). The Stanford Encyclopaedia of Philosophy devotes an entire chapter to the non-trivial question in the heading (Frigg and Hartmann, 2020). For the purposes of this chapter, a useful definition is: "*Models are finalised cognitive constructs of finite complexity that idealise an infinitely complex portion of reality through idealisations that contribute to the achievement of a knowledge on that portion of reality that is reliable, verifiable, objective, and shareable*" (M Viceconti, 2011). Models are a way we humans think about the world. In science, models idealise a quantum of reality:

- To memorise and logically manipulate quanta of reality (*Descriptive models*)
- To combine our beliefs on different quanta of reality in a coherent and non-contradictory way toward the progressive construction of a shared vision of the world (*Integrative models*)
- To establish causal and quantitative relationships between quanta of reality (*Predictive models*)

Predictive models are used in science primarily for two purposes:

- as tools used in the development and testing of new theories
- as tools for problem-solving

In this second use purpose, we define the *credibility* of a predictive model as its ability to predict causal and quantitative relationships between quantities in the natural phenomenon being modelled, as measured experimentally. Thus, the first foundational aspect of a model's credibility is the complex relationships that predictive models have with controlled experiments.

3.3. A short reflection on the theoretical limits of models and experiments

Nature is infinitely complex and its mere observation, while useful to formulate explanatory hypotheses of why a certain phenomenon occurs, is not sufficient to test whether such hypotheses are true. To attempt the falsifiability of an explanatory hypothesis, we need real-world observations or a *controlled experiment*, or experiment for short. In an experiment, we intentionally perturb the system under investigation and observe how it responds to this perturbation. By controlling some of the

variables that describe the system's state and observing how other state variables change, we can reject all hypotheses that are inconsistent with the results; the hypothesis that resists all our falsifiability attempts is tentatively assumed to be true.

Controlled experiments are extremely challenging in life science because of the complexity and entanglement of living organisms. The most realistic experiment is the one where we merely observe the system, but even in that case, because of the observer effect, by the simple act of observing the system, we perturb it; then, human beings cannot achieve a hundred percent (100%) realism. As soon as we perturb the system of interest, what we observe is not the system per se but an *experimental model* of it. In other words, even an observational study is a model of reality. As soon as we investigate reality with a model (which we believe is always the case), the key question is the "*Degree of Analogy"* between the model and the reality being modelled: How close does the model capture the functional aspects of the reality that we are trying to understand? It might look completely different, but if it works like the portion of reality under investigation, it is a good model.

A big advantage of experimental models is that their Degree of Analogy with the reality they model can be inferred from how they were built. Every experimental model contains a fraction of physical reality. The bigger this fraction, the higher the Degree of Analogy of the experimental model.

Too frequently in medicine, we confuse analogy with *homology*: Two biological systems are homologous if they have evolved from the same origin or from a common ancestor, regardless of their function. Then, we consider mice as experimental models of humans because both are terrestrial vertebrates with common ancestors. Still, concerning a specific physiological function, a mouse might be farther from a human than a fruit fly.

However, there is unquestionably a relationship between analogy and homology. The closer our experimental model is to the reality we want to investigate, the more likely the model will have a strong analogy with such reality. Therefore, even if it is done because of homology and not of analogy, in general, a randomised clinical trial of a new drug is more analogous to the reality of the use of that drug in clinical practice than an animal study on the efficacy of that drug, which in turn is more analogous than an *in vitro* experiment in cell culture. This might not always be the case, but it frequently is.

Thus, we can infer the Degree of Analogy an experimental model has with the reality we are investigating by looking at how the experiment was built. The more controlled the experiment, the heavier perturbation we make to the physical reality and the lower the degree of analogy. So experimental models trade off controllability with the degree of analogy, which can be inferred from how the experiment was built.

It should be noted here that the controllability of an experiment in the context of life science is not only limited by the trade-off with the Degree of Analogy. Living organisms are very complex and highly entangled, which means that perturbing one specific aspect may (and usually does) change many other aspects, sometimes in fairly unpredictable ways. To this, we need to add all the ethical limits of animal and human experimentation. Sometimes the optimal experimental design is not possible for ethical reasons.

There is another way to build models of reality. As introduced above, models can be defined as "*finalised cognitive constructs of finite complexity that idealise an infinitely complex portion of reality through idealisations that contribute to the achievement of knowledge on that portion of reality that is objective, shareable, reliable and verifiable*" (M. Viceconti, 2011). If we accept this definition, models can be built not only by perturbing/manipulating the physical reality we want to investigate

(experimental models) but also by any other type of idealisation process. Here, we are interested in "*in silico*" models built through computational modelling and simulation of specific idealisation processes.

The idealisation processes we use to build silico models can differ greatly. Statistical inference models are built through inductive reasoning framed in a frequentist or Bayesian theory of probability; biophysical mechanistic models are built by deductive reasoning starting from tentative knowledge that has resisted extensive attempts of falsifiability (laws of physics). While these differences will become vital in other chapters, here it will suffice to recognise that *in silico* models are built through some idealisation process.

We notice two significant differences if we compare in silico and experimental models. The first is that the Degree of Analogy an in silico model has with the reality under investigation cannot be inferred by how we built the model. Since there is no grounding with the physical reality typical of experimental models, **the degree of analogy must be demonstrated for each *in silico* model**.

This is a major shortcoming of *in silico* models, which would almost always make us prefer experimental models if not for another important difference: In *in silico* models, the controllability is entirely independent of the Degree of Analogy. This means that we could, in principle, consider the use of *in silico* models to reduce, refine, and replace experimental models when it is possible to demonstrate their Degree of Analogy with the reality being modelled and when that Degree of Analogy is higher than that offered by experimental models with similar levels of control. The second motivation for using *in silico* models to reduce, refine and replace experimental models is when for the same Degree of Analogy and the same level of controllability, *in silico* models can provide the required answer faster and/or at a lower cost. A third motivation comes from the observation that even for experimental studies within the currently accepted ethical boundaries, every animal and human experiment has an *ethical cost* that should be minimised as much as possible.

We can infer the Degree of Analogy of experimental models simply by how they are built; all we need to do is to quantify their validity and reliability. On the contrary, with *in silico* models, we must demonstrate that the model has the necessary Degree of Analogy for each Context of Use before we can use it to reduce, refine, or replace experimental models.

3.4. Model for hypothesis testing, models for problem-solving

In the previous section, we introduced experimental models as a necessity of the scientific method, which requires that each hypothesis born out of the observation of a natural phenomenon is relentlessly challenged with controlled experiments designed to falsify this hypothesis. This is the classic use of models in fundamental science when the goal is to increase our knowledge of the world around us. But there is another use for models, whether experimental or *in silico*: Problem-solving. In his famous book "All Life is Problem Solving" (Popper, 1994), Karl Popper insists on using tentative scientific knowledge to solve problems affecting human life, including healthcare.

All our reflections in this position paper are related to the use of models for problem-solving, and in particular, to a specific class of problem: Determining, before its widespread use, if a new medical product is sufficiently safe and effective to justify its marketing authorisation, by allowing patients to be the ultimate users of such new medical product.

While in knowledge discovery, the focus is on the falsifiability of hypotheses, in problem-solving, we assume that the knowledge used to build our predictive models (if any) is tentatively true.

However, this does not automatically imply that the model predictions will be accurate; several factors, which we will detail in the next section, may introduce errors in the prediction. Therefore, it is necessary to systematically assess its Degree of Analogy before a predictive model is used in a mission-critical context (e.g., a predictive model of a medical device or medicine that may save a patient's life).

Another related dichotomy frequently used to separate statistical models from machine learning models is between *inference* and *prediction*. Inference aims to generalise for an entire population the properties observed in a sample of such a population. The purpose of inference models is *representational* in nature. Prediction aims to forecast unobserved data, such as future behaviour (e.g., in the business context, predictive modelling uses known results to create, process, and validate a model that may be used to forecast future outcomes in a specific context of use). The purpose of predictive models is *predictive* in nature. While inference is backed by a robust mathematical theory (probability theory) and, in particular, by the Law of Large Numbers, which resisted extensive falsifiability attempts, this theory does not necessarily apply to data-driven predictive models, which makes the evaluation of the Degree of Analogy for data-driven models epistemologically challenging.

3.5. Assessing the Degree of Analogy of a model: evidence by induction

The predictive accuracy of a model can be estimated by comparing its predictions to the results of a matching controlled experiment. Matching here means that the model should be informed with a set of inputs that quantify the independent variables of the controlled experiment, the quantities we control in the experiment. By doing so, we assume that the model is a model belonging to that specific experiment. Thus, the predictive accuracy (for that particular set of inputs) is the degree of agreement between the values of the dependent variables measured in the controlled experiment and the same values as predicted by the model. This activity is usually called *experimental validation of a predictive model*. It should be noted that for classic validation studies, it is expected that the errors affecting the measurement methods used in the experiment to be negligible if compared to those affecting the model's prediction; this allows the assumption that the measured value is "true" and the difference between prediction and measurement is due to the errors affecting the model. When this is not the case, comparing the model to the experiment becomes much more complex.

There is a major issue with this approach: its inductive nature. By validating the model with one experiment, we estimate its predictive accuracy for those input values. This only allows us to say that the model has a certain accuracy when used to predict a specific condition described by those input values. *A priori*, nothing can be said about the model accuracy for other input values. Of course, it can be done another validation experiment and calculating the model's predictive accuracy for a second input set. Still, again, this will extend our validity statements only to this second condition. We can do many validation experiments and try to build by induction a general validity for our model, or we can look at the nature of the predictive error the model being tested exhibits and find patterns and regularities.

The analysis of how the prediction error is composed is more commonly used in the validation of mechanistic, knowledge-based predictive models. In contrast, the validation by induction is typical for data-driven predictive models. The separation of the predictive error in its numerical, epistemic, and aleatoric components is the central motivation for the so-called Verification, Validation, and Uncertainty Quantification ('VVUQ') (Viceconti et al., 2020b).

3.6. The theoretical framing of VVUQ

VVUQ developed within engineering sciences as an empirical practice without clear theoretical foundations. This may sound surprising, but historically also the most important numerical methods in engineering, like finite element analysis, were first developed as empirical methods and only later found a theoretical framing as a special case of the Galerkin method. Like all practices, different practitioners have different interpretations of what VVUQ means. Also, VVUQ is frequently used in engineering science without many questions on why such a process should inform us better about the credibility of knowledge-based predictive models than any other approach.

However, for the purpose of this chapter, it is important to make explicit the theoretical framing that supports the use of VVUQ. This is because, as we will see, this approach relies on several assumptions, which might not always be true when the evaluated model predicts complex living processes. Here, we provide a summary; full details can be found in (Viceconti et al., 2020b).

There are three possible sources of predictive error in a knowledge-based model:

- The *numerical error* we commit by solving the model's equations numerically;
- the *epistemic error* that we commit due to our incomplete, idealised, or partially fallacious knowledge of the phenomenon being modelled;
- the *aleatoric error* due to the propagation of the measurement errors that affect all our model inputs.

If we compare a model's prediction to the result of a controlled experiment, we will observe a difference caused by all these errors. The VVUQ process aims to separate these three components of the predictive error. If the model is solved appropriately, we expect the numerical error to be negligible compared to the other two. We expect the aleatoric error to be comparable to the measurement errors that affect the model's inputs. If this is not the case, the model might have mathematical or numerical instabilities. In other words, we want to be reassured that the epistemic error is the predominant component of the predictive error.

Verification activities aim to quantify the numerical error. At the risk of oversimplifying, verification tests the model with special input values where epistemic and aleatoric errors are exactly null or asymptotically convergent to null. While the verification is performed for these special input values, because it is generally true that numerical errors are independent of or only weakly dependent on the inputs, we assume that the numerical errors found with those special input values will remain roughly the same for any other input value.

Uncertainty quantification explores how the experimental errors affecting the model inputs propagate within the model and affect the predicted values. The input values are perturbed according to the probability distribution of the experimental error affecting them, and the variance induced in the predicted outputs is recorded. Uncertainty quantification directly estimates the aleatoric error for a specific set of input values. It is usually assumed that how the error due to the inputs' uncertainties propagates into the model's predictions is independent of the specific values of the inputs.

Validation activities rely on two assumptions. First, the numerical errors are negligible compared to the other two sources of error. Second, the aleatoric error is normally distributed around a null mean. If this second assumption is true, the effect of the numerical errors will be negligible when we calculate the predictive error as an average (e.g., root mean square average) over multiple experiments. In this case, the aleatoric errors will also net out, leaving the average predictive error as a good estimate of the epistemic error.

The last step in the VVUQ is the so-called ***applicability analysis***. While we tend to assume that numerical and aleatoric errors do not depend on the specific values of the inputs, such an assumption cannot generally be made for the epistemic error. On the contrary, it is expected that any idealisation holds within a limit of validity, and as we get closer to those limits, the epistemic error will increase. There are various approaches to evaluating the applicability of a model. Still, most rely on one fundamental assumption: if two input sets are similar, the two output sets will also be similar. Suppose the model tends to show similar epistemic errors for all tested inputs. In that case, we can consider that for all other input values within the range of values tested during the validation, the epistemic error will also be similar. The further the model is used in terms of inputs from the range of values tested in the validation, the lower the reliability of the estimate of epistemic error we obtained with the validation activities. Another issue to consider, as mentioned above, is that every mechanistic model relies on theories, and every theory has some limits of validity. Once the inputs get closer to such limits, the model's predictive accuracy can degrade considerably.

3.7. Levels of credibility testing

The combination of VVUQ and applicability analysis extends the concept of the model's credibility to combinations of input values that have not been experimentally validated. However, the issue of assessing if a predictive model is credible enough for a specific context of use has two additional aspects: the level of credibility at which we test the model and the minimum predictive accuracy below which we must reject the use of the model. The level of credibility testing is not an attribute of the model; it is the expectation of predictive accuracy for that model, which we define by choosing against what we calculate the predictive accuracy of the model. There are three possible levels (fig 3.1):

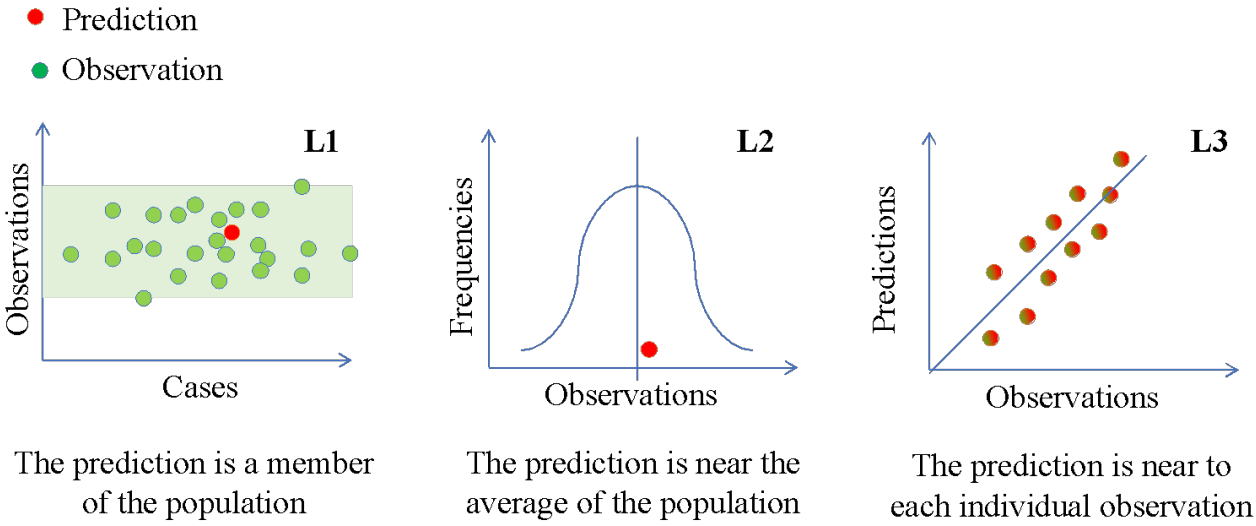


Figure 3.1. Definition of levels of credibility for a predictive model

- At the lowest level of credibility testing (L1), models aim to predict a value within the range of values observed experimentally over a population. Here, the predictive accuracy is measured in terms of the probability that the predicted value for each Quantity of Interest ('QoI') is a member of the population of values measured experimentally.

- The second level of credibility testing (L2) expects the model to predict some central properties accurately (e.g., the average) of the distribution of values observed experimentally over the population. Here, the predictive accuracy is quantified by measuring the distance for each QoI between the predicted value and the average of the values measured experimentally.
- Lastly, the highest level of credibility testing (L3) expects the model to accurately predict the value observed for each member of the population. Here, the predictive accuracy is calculated as a p-norm of the vector of differences between the predicted value and the measured value for each member of the population. Most commonly, a 2-norm is used (root mean square error), but a more restrictive infinity-norm, where the measure of the error is the maximum error found among all members of the population, may also be used.

While this taxonomy of the level of credibility testing is not considered in any current regulatory document, we recommend it be considered in future guidelines and standards.

3.8. The conundrum of validating data-driven models

The whole framework of the model's credibility based on VVUQ plus applicability was developed, having in mind models built starting from a causal explanation of the phenomenon being modelled (mechanistic models). By considering epistemic errors, VVUQ-based credibility accepts that the prior knowledge we use to build the model might be inaccurate, but it is always present. And in most cases, such knowledge is expressed with mathematical forms whose properties summarise such knowledge. For example, all theories expressed with differential equations implicitly assume that the variation of the quantities of interest over space and/or time occurs smoothly. This, in turn, derives from an essential physical knowledge of the conservation of mass, momentum, and energy. In fact, many of the implicit assumptions that the use of VVUQ to assess a model's credibility that we listed in the previous sections are usually valid under such assumptions.

But this raises an important question: can credibility assessment based on VVUQ plus applicability be used also for models that are not built with some prior knowledge (hereinafter referred to as 'data-driven models')? The short answer is no; here, we provide some theoretical justifications for this conclusion.

In probability theory, if we are sampling some population properties, the Central Limit Theorem ('CLT') tells us that such sampling will eventually converge to a normal distribution. The Law of Large Numbers ('LLN') states that with enough samples, the estimates of certain properties of the probability distribution, such as average or variance, will asymptotically converge to the true values for that population. This guarantee of asymptotic convergence makes it possible to infer the properties of a distribution from a large but finite number of samples.

Let us now consider the use of a statistical model as a predictor. Here, using statistical inference, it can be built the hypothesis that the value of the dependent variable Y can be predicted given the values of a set of independent variables $[X_1, .., X_n]$ so that $\underline{Y} = f(\underline{X}_{1, ..., } \underline{X}_n)$. Here for simplicity of treatment, we assume that the variables X_i can be quantified without any uncertainty. By inferring the relations and correlations between $\underline{X}_1$ and $\underline{Y}$, it can be built an estimate of $f()$, (called $f'()$), which it can be used to predict $\underline{Y}$ for combinations of $\underline{X}_1$ that have not yet been observed experimentally. If the LLN theorem holds, it is sufficient to have a finite number of observations $[\underline{X}_i, \underline{Y}_i]$ to build $f'()$.

But if we now want to quantify the predictive accuracy by comparing the value predicted $\underline{Y}'(\underline{X})=f'(\underline{X})$ to that observed experimentally ($\underline{Y}$) for a finite number of $\underline{X}_i$ sets, does the LLN still apply? Given a

large enough set of validation experiments where it is observed $P(\underline{Y} \mid \underline{X_i})$, is there a theoretical foundation to assume that the estimate of the average prediction error $e' = ave(\underline{Y} \mid \underline{X_i})$ tends to the true value e it would get if we could validate the predictor with an infinite number of experiments? Does the average prediction error estimate tend asymptotically to the true average prediction error?

When estimating, we learn about the characteristics of a population by taking a sample and measuring those characteristics. The fact that we have a sample brings about variability (uncertainty), normally described by a probability distribution whose parameters are related to the characteristic of interest. Usually, the more information we have about the characteristic (the larger the sample size), the larger the accuracy (estimating the correct value of the characteristic) and precision (decreasing the uncertainty) of the estimation. If some very mild conditions apply, we can assume the variability in the estimators follows a normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where x is the measured quantity, μ is the mean, and σ is the standard deviation.

Now, if the objective is to predict, we must consider an extra source of uncertainty. This changes the nature of the statistical problem, indeed. The formal problem is to analyse $P(\underline{Y} \mid \underline{X})$, with $\underline{Y}$ the characteristic of interest and $\underline{X}$ all the relevant data available. Of course, $\underline{Y}$ and $\underline{X}$ are related formally by a set of parameters relevant to $\underline{X}$ and $\underline{Y}$; for simplicity, formalise them as $g(\underline{X} \mid \underline{A})$ and $g(\underline{Y} \mid \underline{B})$, with $\underline{B} = B(\underline{A})$ an invertible function. So, you learn about $\underline{A}$ — hence $\underline{B}$ — from $\underline{X}$ and that knowledge is described by $f(\underline{B} \mid \underline{X})$ — which can be Gaussian and with increasing precision and accuracy as above. We can then use this knowledge to inform $g(\underline{Y} \mid \underline{B})$, but even if you knew $\underline{B}$ exactly, there is still a source of uncertainty in $g(\underline{Y} \mid \underline{B})$ that cannot be reduced further; moreover, the shape of g is not warranted to be normal at all.

There are alternative ways to include the information about $\underline{B}$ in $g(\underline{Y} \mid \underline{B})$. But in any case, you can grasp what is going on formally by using the law of iterated expectations. The expected value of $\underline{Y} \mid \underline{X}$ can be calculated as the expected value of the expected value of $\underline{Y} \mid \underline{X},\underline{B}$. The more you learn about $\underline{B}$ from $\underline{X}$, the better the estimation of the mean of $\underline{Y}$. So, larger sample sizes yield more accurate predictions. However, this is not necessarily the case for the precision of the prediction. The variance of $\underline{Y} \mid \underline{X}$ can be calculated as the expected value of the variance of $\underline{Y} \mid \underline{X},\underline{B}$ plus the variance of the expected value of $\underline{Y} \mid \underline{X},\underline{B}$. The second term decreases with the sample size, but the first one does not and depends on the distribution of $\underline{Y} \mid \underline{B}$. The validation of a predictive model must take both sources into account.

Let us assume we are interested in predicting a quantity $\underline{Y}$, which depends on a set of values $\underline{X}$. $f()$ is a predictive model that provides an estimate of $\underline{Y}$, which we call $\underline{Y}'$. The concept of model credibility assessment based on VVUQ is that the model $f()$ is mechanistically defined, so we know that $\underline{Y} = f(\underline{X})$, and any other variable outside of the set $\underline{X}$ has little or no effect on $\underline{Y}$ (or the effect is mediated through $\underline{X}$).

An important implication of all this is that for data-driven models, the smoothness of the prediction error is not guaranteed, as it is for mechanistic models. In mechanistic models, we can assume that our error $e = \underline{Y}\text{-}\underline{Y}'$ depends only on $\underline{X}$, so if we test $f()$ for $\underline{X_1}$ and $\underline{X_2}$, where $\underline{X_1} \approx \underline{X_2}$, the prediction error will be similar, $e_1 \approx e_2$. This also means that if $e(\underline{X_1})$ is the prediction error for $\underline{X_1}$, and $e(\underline{X_2})$ is for $\underline{X_2}$, $e(\underline{X_i})$ will be close to e_1 and e_2, if $\underline{X_i}$ is close to $\underline{X_1}$ and $\underline{X_2}$. In other words, if the model is validated for a range of X_i, it could safely be assumed that the error will be similar for any other $\underline{X}$ close to $\underline{X_i}$. But this cannot be said for data-driven models, such as Machine Learning (ML) models

because there is no guarantee that $\underline{Y}$ is a function only of $\underline{X}$. We cannot, as we do with mechanistic models, test the ML model for a finite number of cases, and assume that average accuracy will stay for any other case close to those tested. This is pure induction: by testing the ML model against ten experiments, it can only be said that the error with those ten cases is that, but the next could show an error totally different, even for a $\underline{X}_i$ close to the ten we already tested.

This poses two significant problems when the VVUQ approach is used to assess the credibility of data-driven models. The first is that while in mechanistic models, the variance of $\underline{Y}$ can be mainly explained with the variance of $\underline{X}$, in data-driven models, this is not assured. As explained above, X may include variables that have little effect on $\underline{Y}$. Also, we have no guarantee that all variables affecting $\underline{Y}$ are included in $\underline{X}$. This uncertainty is the primary cause of the so-called '*concept drift*', which sometimes causes a data-driven model to perform much worse than it did on the training set when the test is done against an independent validation set.

The second is the lack of smoothness in the prediction error. As explained in Chapter 3.4, the applicability analysis presumes that the model's prediction error will vary smoothly as the inputs of the model are varied. This makes it possible to assume that if the model is used with inputs "near" to the values for which the model has been validated against experimental results, the error affecting the prediction will be similar to that quantified with the validation. However, such assumptions cannot be made for data-driven models.

The "*Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan*" that FDA published on January 2021[8] explicitly refers to introducing a so-called Predetermined Change Control Plan in the US regulatory system. Thus, a total product lifecycle ('TPLC') regulatory approach to AI/ML-based SaMD, is designed considering the iterative, adaptive, and autonomous natures of AI/ML technologies. Essentially, the idea is that the validation of data-driven models is a continuous process where we continuously extend the test set, re-evaluate the model's predictive accuracy, and regenerate it, using this test set as an extended training set.

Our reflections would suggest that this approach is not only possible but should be the only acceptable approach. In light of the discussion above, the idea of a "frozen" data-driven model, validated with the same VVUQ used for mechanistic models, seems unwise. This conflicts with the obvious need for regulatory approval processes to base the decision on a prediction made using a "frozen" model.

3.9. Conclusions

Some conclusions can be drawn that can inform the rest of this position paper.

The human mind can investigate reality only through cognitive artefacts we call models. Whether we use a mathematical model or a controlled experiment (including observational studies), we are always dealing with models of reality; ultimately, what matters is the Degree of Analogy that the model has with the reality being modelled. The main advantage that experimental methods have over *in silico* methods is that in experimental models, the Degree of Analogy can be easily inferred by their design, whereas for *in silico* methods the Degree of Analogy must be assessed case by case.

When a model is used to make predictions in the context of problem-solving, the Degree of Analogy with the reality being modelled becomes the *credibility* of the model's predictions. In general, credibility can be assessed only by induction; then, if we quantified the predictive accuracy of our

[8] https://www.fda.gov/media/145022/download

model against hundred (100) experimental measurements, we could only state the credibility of the model in those 100 experimental conditions. The number of experimental conditions for which the predictive accuracy needs to be tested, called the "*solution space*", is infinite (∞). However, under certain assumptions, we can analyse how the various components of the predictive error (numerical, aleatoric and epistemic) vary over the solution space using a process known as VVUQ plus applicability analysis. This makes it possible to estimate the predictive accuracy over the entire solution space with only a finite number of validation experiments.

The assumptions that make the VVUQ process possible are usually valid only if the model being tested is built with some degree of prior knowledge (mechanistic model). This is not the case for data-driven models, which can be tested only by induction.

3.10. Essential Good Simulation Practice recommendations

- The human mind can understand reality only through *models*. Models are finalised cognitive constructs of finite complexity that idealise an infinitely complex portion of reality. Their usefulness is measured by their ability to capture the functional aspects of interest of the portion of reality that we are investigating. This measure is called the Degree of Analogy.
- In each portion of reality, the functional aspects of interest can be observed experimentally or predicted through inductive or deductive reasoning. All these methods of investigation are models. However, the Degree of Analogy of experimental models can be directly inferred, whereas that of predictive models must be demonstrated by comparisons with controlled experiments. In other words, experiments are not necessarily more trustworthy than predictions, but their trustworthiness is easier to assess.
- Predictive models can be divided into prevalently data-driven models and prevalently mechanistic models. In prevalently mechanistic models, the Degree of Analogy can be established by decomposing the predictive errors in numerical, aleatoric, and epistemic errors through a process known as Verification, Validation, and Uncertainty Quantification. But in prevalently data-driven models, the Degree of Analogy can only be estimated by induction, using a total product lifecycle regulatory approach.

4. MODEL DEVELOPMENT

Chapter editors:

- Alexander Kulesza (Novadiscovery)
- Axel Loewe (Karlsruhe Institute of Technology)

Contributors (random order):

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Steve Kreuzer (Exponent)
- Eulalie Courcelles (Novadiscovery)
- Fianne Sips (InSilicoTrials Technologies)
- Chiara Nicolò (InSilicoTrials Technologies)
- Enrique Morales Orcajo (Ambu Innovation GmbH)
- Giulia Russo (University of Catania)
- Francesco Pappalardo (University of Catania)
- Raphaëlle Lesage (VPH institute)
- Andrea Stenti (Université libre de Bruxelles ULB)
- Marc Horner, Ansys Corp.
- Martha De Cunha Maluf-Burgman (Edwards Lifesciences)

4.1. A risk-based paradigm of model development as a function of its Context of Use

Good Simulation Practice implies that a computational model considered for a simulation task has also been developed according to good practice. In this Chapter, an attempt is made to summarise and synthesise good practices in computational model development. Considering numerous different model types (a recent report of the US Food and Drug Administration (FDA)[9] mentions 39 different modelling classes), a high level of abstraction is needed. Therefore, this Chapter focuses on the model development and implementation as a process rather than concrete model-type specific recommendations. Generic model definition and design recommendations are addressed in section 4.2.2.

Whether one develops the predictive model from scratch or from existing libraries and solvers, computational model development shares many commonalities with software development:

a) Models transform user inputs into outputs.

b) Models can be developed as standalone units or part of larger systems/platforms.

c) A model's life cycle is similar to that of software.

d) The concrete implementation of a predictive model is often part of a software.

This considered, it is reasonable to explore existing standards for Software Life Cycle (SLC) management (systems and software engineering) as a starting point for good practices in model development, defined according to widely agreed-upon "best ways of doing" it[10] - relevant for

[9] https://www.fda.gov/media/163156/download

[10] "ISO standards are internationally agreed by experts." https://www.iso.org/standards.html. Accessed 19 Sept. 2021.

application and uptake in mission-critical and highly regulated environments. Many different programming languages and software development paradigms exist and can be used for developing computational models under consideration of process-level software development good practice - agnostic of procedural and content-related details. We will therefore leverage the similarity of computational models and software and map good software development onto the former.

Model developers must acknowledge that their "product" may operate under a "regulated" environment and that regulators will perform a benefit-risk assessment. Any regulatory effort in a mission-critical domain faces the challenge of balancing the need for the lowest possible level of risk for the patient and the economic viability of product development, without which no product could be brought to the market. Risk assessment, clinical evaluation, and validation revolve around the "intended purpose" of the Medical Device Regulation (MDR). There is a debate around how risk management should be implemented in the development of the device: following the concept of risk "As Low As Reasonably Practicable" (ALARP) as proposed by the ISO 14971:2019 or risk reduced "As Far As Possible" (AFAP) as requested by the new EU regulations[11]. Concrete regulatory recommendations on transferring these concepts to computational models do not exist yet except for an FDA draft guidance[12] document. We, therefore, adopt a risk-based approach for the development of computational models, where the level of scrutiny (in terms of model life cycle management) is proportional to the risk (assessors would ascertain) that a predictive model can pose (e.g., for patients) according to pre-defined CoU(s).

This Chapter, thus, focuses on the application of a risk-based approach (Figure 4.1) for the process of model and simulation software development: good practice establishes a minimal set of process-level requirements (Figure 4.1) for all models (even low/medium-risk ones) while it requires more comprehensive compliance with relevant industry standards for medium to high risk, critical applications and models with substantial impact on regulatory decision-making. Here we focus on developing modelling and simulation software, whereas in Chapter 5, the focus is on the result of this process, the actual implementation (the model).

In the following, we first highlight some industry standards and a potential mapping to elements of model development. We then iterate through the stages of a life cycle model relevant to model development. In each section, we adopt a viewpoint from low and high model risk to deriving two (extreme) levels of compliance with the cited industry standards and good simulation practices for model development (Tables 4.2-4.6).

It should be stressed that this chapter's scope is limited to model development practices and does not consider model use and validation aspects, which are covered in Chapter 5.

[11] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745

[12] https://www.fda.gov/media/154985/download

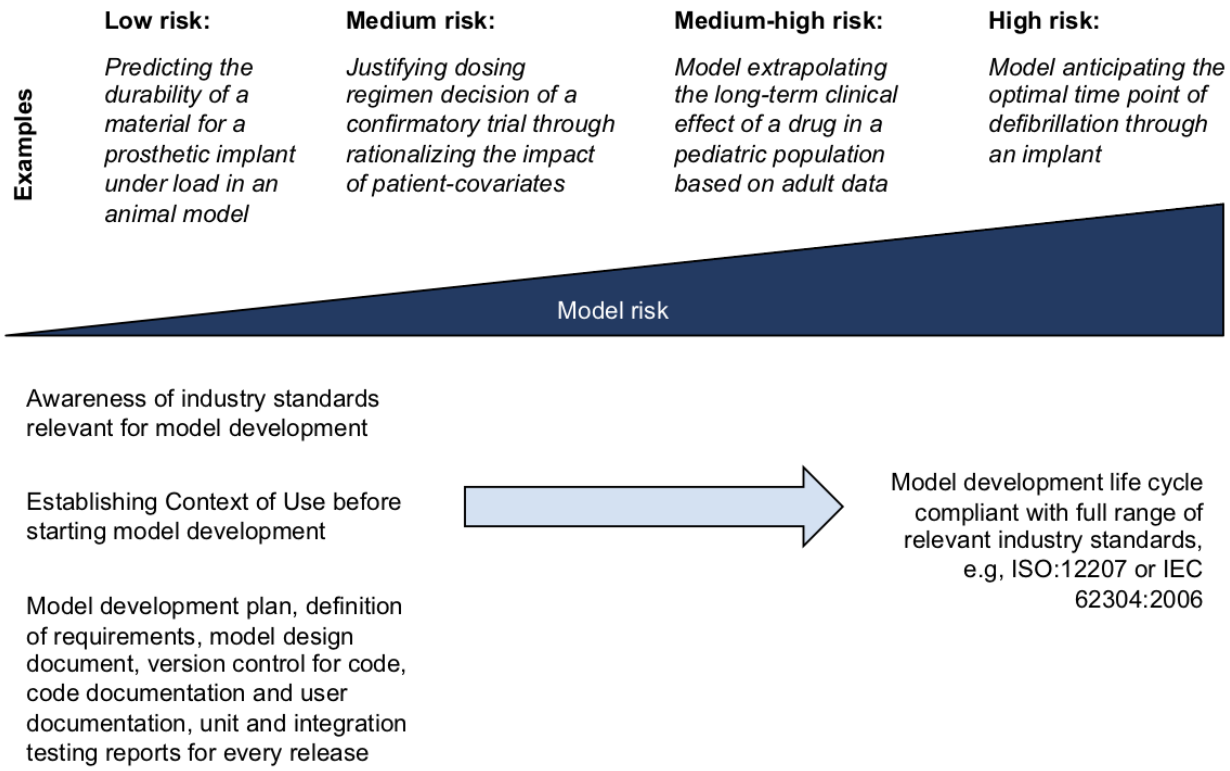


Figure 4.1. A risk-based approach for model development planning (simplified; for a complete model risk assessment, see ASME VV-40:2018). While a lean plan can suffice for low-risk CoUs, higher-risk projects must gradually consider the full range of relevant industry standards detailed in section 4.2. Note that during development, the model risk needs to be anticipated by the developer.

4.2. SLC industry standards and relevance for model development

Above, we introduced a risk-based approach for model development, implementing different levels of compliance with industry standards. As no industry standard for computational models yet exists, the next best option is to adopt and apply standards and best practices from related areas similar to computational model development. Of particular interest to the development of predictive models is the great body of process-level knowledge and recommendations available for software development – not only because of the analogy between software and model development but also because the developed model and the software in which it is implemented are often intertwined.

For software development, two existing standards are relevant to the model development process: ISO/IEC/IEEE:12207[13] and ISO/IEC:62304[14]. The former applies to every software package or system, whereas the latter is specific to medical device software. Other relevant best practice documents include the National Aeronautics and Space Administration (NASA) handbook on model development (NASA-STD-7009A[15] and NASA-HDBK-7009A[16]). A full mapping of all the model

[13] https://www.iso.org/standard/63712.html
[14] https://www.iso.org/standard/38421.html
[15] https://standards.nasa.gov/standard/NASA/NASA-STD-7009
[16] https://standards.nasa.gov/standard/nasa/nasa-hdbk-7009

development activities, with processes from all potentially applicable industry standards, is out of this chapter's scope. Instead, we intend to highlight the opportunities where an explicit consideration of industry standards can help to ensure overall quality, especially in critical applications.

- **The ISO/IEC/IEEE 12207:2017 standard "Software life cycle processes"** covers different process groups, including I) Agreement, II) Organisational project-enabling processes, III) Technical management processes, IV) Technical processes. Process groups III and IV are most relevant for a given model development project in a given organisational structure (for definitions of a process, see, e.g., ISO/IEC/IEEE 24774:2021). We stress that management and technical processes are orthogonal.
    - **Technical management processes** are concerned with managing the resources and the assets allocated by the organisation's management and applying them. Technical management comprises a) Project planning, b) Project assessment and control process, c) Decision management process, d) Risk management process, e) Configuration management, f) Information management, g) Measurement, h) Quality assurance process.
    - **Technical processes** are concerned with technical actions throughout the life cycle. Technical processes transform the needs of stakeholders into a product or service. Technical processes include a) Mission analysis, b) Stakeholder needs and requirements definition, c) Systems/software requirements definition, d) Architecture definition, e) Design definition, f) System analysis, g) Implementation, h) Integration, i) Verification, j) Transition, k) Validation, l) Operation, m) Maintenance and n) Disposal, of which (a-h) may be classified as "development" and are covered in detail in this Chapter while (i-k) are more relevant for model validation (see Chapter 5). To qualify for sustained use or regulatory approval, also (l-n) need to be considered, where maintenance is often challenging in a research setting (Anzt et al., 2021).
    - ISO/IEC/IEEE 15288:2015 establishes a common framework of process descriptions for the life cycle of systems and is often used in conjunction with ISO/IEC/IEEE 12207:2017. Additionally, ISO/IEC/IEEE 24748-3 describes the application of ISO/IEC/IEEE 12207:2017.
- **The standard ISO/IEC 62304:2006 "medical device software – software life cycle processes"** may be more straightforwardly adopted as it is targeting regulated environments in healthcare and is recognized by both the European Union as a Harmonised Standard and the United States FDA as a Recognized Consensus Standard. The structure of this standard is similar to ISO/IEC/IEEE 12207 (see overlap indicated in Table 4.1). In fact, modern development platforms that combine the ability to develop, secure, and operate software, can be operated to establish compliance (see here[17] for the case of ISO/IEC 62304).

These ISO/IEC standards do not prescribe any particular life cycle model, acknowledging that software development processes should be oriented towards the project objectives. Instead, it defines a set of processes, termed life cycle processes, which can be used in the definition of the SLC. Depending on the SLC model used, the development phases may be classified as follows:

- **Analysis & Requirements:** consideration of the real-world system and the possibilities of what M&S can do for it; derive requirements from the mission analysis, the user perspective and the (integrated) system viewpoint.

[17] https://about.gitlab.com/solutions/iec-62304/

- **Design**: collection of information and definition of concepts to include in the proposed model; iterative process of creating the detailed, verifiable, and validated specification of the model and simulations for an intended use.
- **Implementation & Integration:** realisation of the technical implementation of the model design in line with the requirements, specifications, and intended use.
- **Testing**: checking to determine if the model meets all requirements and operational intentions.
- **Maintenance**: release of the software, archiving of artefacts, life cycle management.

Table 4.1. Relation between phases outlined in this document and most relevant phases and processes in selected industry standards and best practice documents on (model development) planning.

SLC phase	ISO IEC 12207	ISO 62304	NASA-HDBK-7009A
Analysis	Mission analysis	Software development planning	Model initiation
Requirements	Stakeholder needs and requirements definition	Software requirements analysis	
	Systems/software requirements definition		
Design	Architecture definition	Software architectural design	Model concept development
	Design definition	Software detailed design	Model design
	System analysis		
Implementation	Implementation	Software unit implementation and verification	Model construction
	Integration	Software integration and integration testing	
Testing	Verification (developer and end user)	Software system testing	Model testing and release
	Transition		
	Validation (Developer and end user)		
Maintenance	Operation (end user)	Software release	Model use (end user)
	Maintenance		Model and Analysis Archiving (developer and end user)
	Disposal		

Orthogonal to the technical processes, adherence to management processes can be beneficial or required. Also, adherence to industry standards rarely concerns one portion of the set of business processes but more likely impacts a large set of operations. Other standards also apply, which especially cover quality management/assurance, such as ISO 13485 (quality management system for designing and manufacturing medical devices) or, even more generally, **ISO 9001**. Also, consideration of a service management system specified by ISO/IEC 20000-1, an IT asset

management system specified by ISO/IEC 19770 (all parts) and an information security management system specified by ISO/IEC 27000 can be relevant.

4.2.1. Analysis & Requirements

Initially, the development life cycle is initiated by a planning phase where the overall mission, problem, and context are analysed, and the actual requirements are defined. Subsequently, a plan defines how the new development will fulfil the mission. In many life cycle models, this initial phase is called Elicitation. Depending on the model, a development plan that is either more project-oriented or more technically oriented can be the better choice. Table 4.2 lists good model development practices by establishing, considering and documenting a model development plan and the requirements definitions document(s) for both a low and high model risk scenario (Figure 4.1 left and right, respectively). Generally, we regard the high-risk recommendation as the gold standard, while simplified processes and documentation can be acceptable for low-risk.

From a model developer's viewpoint, this phase in the SLC overlaps with the definition of a CoU in ASME V&V-40 (see Table 4.1). The model developer must anticipate the required CoUs that the model will aim for (the CoUs anticipated by the developer might be captured in a Concept of Operations document and use cases). In high-risk contexts, the formulation of the CoUs should thus be embedded with requirements definitions aligned with industry standards from related disciplines until specific ones become available.

Planning for the project is often captured in a Project Management Plan. ISO/IEC/IEEE 16326 provides more detail on project planning. The project planning process aims to produce and coordinate effective, workable plans. This process determines the scope of the project management and technical activities, identifies process outputs, tasks, and deliverables, and establishes schedules for task conduct, including achievement criteria and required resources to accomplish tasks. Project planning is an ongoing process throughout a project with regular plan revisions.

Technical planning for a software system is often captured in a Systems Engineering Management Plan, a Software Engineering Management Plan, or a Software Development Plan (SDP). ISO/IEC/IEEE 24748-5 provides more detail on software engineering technical management planning and includes an annotated outline for an SDP. Notably, ISO 62304-2006 and (Rust et al., 2016) suggest an SDP structure commensurate with regulated environments.

Good practice of model development should establish such a development plan (roughly, see "low risk" or, as precisely as possible, "high risk"; Table 4.2) considering the related guidance on software development.

Another important stage in the initial development life cycle phase of Elicitation is the definition of requirements. Requirements can come from many different sources, for example, user needs, functionality, performance, risk, regulatory, processes, or marketing. As stipulated in Table 4.2 (bottom row), both "low-risk" and "high-risk" models should document the requirements for their development. Explicitly adhering to industry standards might not be required for low-risk models.

Table 4.2. Good practices for the Analysis & Requirements phase of model development on the low and high-risk ends of the risk spectrum (see Figure 4.1, left and right, respectively).

	Low Risk	**High Risk**
Requirement definitions (design prerequisite)	Mission requirements (CoU) Risks User requirements System requirements (if any)	Concept of Operations (ConOps) document System Requirements Document (SRD) Requirements Traceability Matrix (RTM) See IEEE/ISO/IEC 29148-2018
Model development plan (project management)	**Model development plan** similar to a software management plan, see e.g. (The Software Sustainability Institute, 2018) Reference Materials (including knowledge and data sources) Development and life cycle planning	Detailed **model development plan** similar to a software development plan. See ISO 62304-2006 and (Rust et al., 2016)

The requirements definition process captures and transforms stakeholder needs into "well-formed requirements" (suitable as inputs for subsequent model development procedures). A "well-formed requirement" shall possess the following attributes: necessary, appropriate, unambiguous, complete, singular, feasible, verifiable, correct, and conforming. IEEE/ISO/IEC 29148-2018 provides more detail on requirements engineering and requirement processes.

In any case, together with the formulated CoUs (equivalent to mission requirements), the list of user and system requirements should be likewise defined. The entire set of requirements enables a common understanding between stakeholders and provides a reference for verification. They must be validated against real-world needs and be feasible to implement and to check (potentially formulated as part of a System Requirements Document (SRD) or Requirements Traceability Matrix (RTM) in high-risk contexts). This enables users to practically judge whether usage scenarios are within the intended CoUs versus ones that might be technically possible but outside of the CoUs.

4.2.2. Design

The key prerequisite for model design is the definition of the CoU(s) during the requirements specification. In the "design specification," these CoU(s) need to be translated into the architecture and component design of the actual model. The formulation of the model in terms of fundamental mathematical equations and parameters is a key aspect, as introduced in Chapter 3. On the one hand, the model needs to be complex enough to fulfil its CoU(s). On the other hand, unique parameter identifiability and parameter uncertainty (either on a population level during calibration or in a given subject during personalisation depending on the CoU) (Galappaththige et al., 2022; Parvinian et al., 2019), numerical accuracy and required computational effort as wells as the options for verification and validation (Pathmanathan and Gray, 2014; Pathmanathan et al., 2017) need to be considered.

Following the law of parsimony, one should aim for the simplest model that can serve the CoU. The decision-making process and limitations that come with this choice must be documented explicitly (Erdemir et al., 2019). Table 4.3 lists specific aspects to be considered during the Design phase for low and high-risk applications.

Table 4.3. Good practices for the Design phase of model development on the low and high-risk ends of the risk spectrum.

	Low Risk	**High Risk**
Model design	Simple **model design document** **Definition of a conceptual model** Which modelling approach is suitable for the CoU? What level of precision is required? Document the limitations of the model **Definition of the architecture design** Focus on functionality, covered hypotheses and phenomena **Description of the detailed design** Considering model-type specific recommendations Focus on expected sensitivity, identifiability, **User interfaces** Human-machine interfaces User experience Dialog design Presentation of information **More detailed operational scenarios and use cases** For example, as a simulation plan or protocol (Developed together with Sponsor (see Chapter 10))	Comprehensive **model design document** (including the elements from low risk, compliant with relevant design documentation standards see for example[18]) Additionally: **Definition of the architecture design** Additionally (if relevant) describe design decisions related to performance compatibility transferability usability adaptability reliability security maintainability **User interfaces** Describe also measures to avoid (user) errors, avoid misinterpretation, to increase use efficiency (ergonomics) and user satisfaction Consider also Usability Engineering File (ISO 14971, IEC 62366-1) **More detailed operational scenarios and use cases** Document decision process according to standards (e.g., ISO 13485, IEC 62304)
Model validation plan	See Chapter 5 and FDA guidance[19]	

Bäker (Bäker, 2018) gave recommendations for computational simulations relevant to operational scenarios and use cases (to be specified by the developer) as well as the simulations to be performed

[18] https://ntrs.nasa.gov/api/citations/20160011412/downloads/20160011412.pdf

[19] "General Principles of Software Validation; Final Guidance for Industry and FDA Staff", https://www.fda.gov/media/73141/download

by the end user (CoUs). The initial Concept of Operations document (see the previous section) has been updated to consider any limitations by the chosen architecture and further operational details.

As evident from the "low-risk" scenario, comprehensive documentation for the model design is needed, irrespective of the risk. The minimum requirement for all models is thus to justify alignment of the modelling concept (and potentially data flow) with the Context of Use anticipated and to document fundamental design choices (ODE, PDE, agent-based model, etc., the granularity) and associated limitations. In all cases, the architecture design (phenomena and hypotheses, composition, input-output transformations) and the detailed design (equations, parameters, initial/boundary conditions) must be documented and justified. This includes potentially the workflow with which unknown parameters are estimated from target data or other workflows to produce tailored versions of the model (mapping onto a geometry, treatments, outputs of other simulations etc.). For many fields, model-type-specific recommendations exist and should be considered. For example, for pharmacological modelling (Byon et al., 2013; Cucurull-Sanchez et al., 2019; Jean et al., 2021; Ke et al., 2016; Overgaard et al., 2015; Zhao et al., 2012). A comprehensive list of all domains is beyond the scope of this Chapter.

Considering relevant external conditions, the decision to use a specific model architecture should be based on the defined requirements (see the previous section). The essential properties of the architecture are often defined by the required internal and external interfaces to be implemented. The design should use established community standards regarding data formats and application programming interfaces (APIs) whenever possible. One should generally favour simple architectures following the established best practices in software engineering, such as information hiding, loose coupling, high cohesion, separation of concerns and hierarchical decomposition.

Design decisions are documented in the software management plan, potentially already created during the elicitation phase. The documentation of the model design should be similar across all model risks, even though the extent and form may differ. Good practice in the definition and description of the architecture and system design is to use diagrams and (standard) graphical notations, such as the Unified Modelling Language (UML)[20], to help communicate with stakeholders, explore potential designs, validate the architecture of the software, and document decisions. Alongside the input and output data description, detailing how these data enter and leave the system (i.e., interfaces) is also mandatory. Other relevant parts of the documentation (for example, a user manual) can only be generated during/after implementation (see next section).

Notice that the design of the model should anticipate validation and verification activities (either by the developer and/or even the user) to be compatible with them. Therefore, a validation plan should be initiated during the design phase (see Chapter 5).

As can be seen on the right side of Table 4.3, full compliance with industry standards relevant for models can necessitate exhaustive documentation of potential pitfalls, errors, practical and life cycle aspects and may need to follow certain forms.

4.2.3. Implementation & Integration

The purpose of the implementation process is to realise a specified system element. This process transforms requirements, architecture, and design (including interfaces) into actions that create a

[20] Unified Modelling Language v2.5.1 (2017) by the OMG standardisation group: https://www.omg.org/spec/UML/

system element according to the practices of the selected implementation technology. This process results in a system element that satisfies specified system requirements, architecture, and design.

The integration process aims to synthesise a set of system elements into a functional system (product or service) that satisfies system/software requirements, architecture, and design. This process assembles the implemented system elements. Previously defined interfaces are activated to enable the interoperability of the system elements as intended.

Table 4.4. Good practices for the Implementation & Integration phase of model development on the low-risk and high-risk ends of the risk spectrum.

	Low Risk	**High Risk**
Versioned model	Storage of code versions in **version control** systems, e.g., git. **Documentation** of the model (both in the form of the source code and a user manual).	Version control, e.g., in line with ISO 12207, through a **Configuration Management process** for the selection of configuration items to be integrated.

During implementation and integration (Table 4.4), established software engineering best practices should be applied (Rust et al., 2016; Anzt et al., 2021). Fundamental requirements specify the use of a version control system, such as git, and software documentation in the form of the source code and a user manual. Requirements and issues should be tracked and linked using the adequate infrastructure. Specific versions of the implemented model must be assigned unique identifiers (e.g., build number and date). In the case of software as a medical device, persistent unique identifiers are required. Release software versions should be archived with relevant artefacts like documentation, test reports, etc.

Automated tests (see next section) can help increase the implementation efforts' efficiency.

The more regulated the environment and the higher the risk of the CoU, the more important it becomes to standardise approaches strictly. Such standards can, for example, include code style guidelines and conventions.

4.2.4. Testing

Testing serves the purpose of checking whether the model was implemented correctly. It answers the question “Did we build the model right?” as opposed to validation, which addresses “Did we build the right model?” Both aspects are covered in detail in Chapter 5, so we only point out a few specific issues from a model developer's perspective here (Table 4.5).

Various forms of testing can be applied during testing, including:

- Regression tests, which run the model with specified input parameters and compare it against previously computed results.
- Simplified test cases with analytical solutions, which furthermore allows evaluation of the quality of the solution.

- Performance tests can help obtain efficient code but are not the focus of the GSP.

Regardless of the approach, testing should be automated as much as possible in continuous integration pipelines. This will increase adoption and adherence by minimising developer efforts to maintain compliance.

Table 4.5. Good practices for the Testing phase of model development on the low-risk and high-risk ends of the risk spectrum.

	Low Risk	**High Risk**
Tests	Automatic tests during development: integration tests system tests / regression tests	Additional automatic tests throughout the life cycle: unit tests consider static analysis maximize unit testing code coverage
User Feedback	optional	Essential

4.2.5. Maintenance

The need for maintenance can arise from multiple causes other than model bugs, such as version updates of the model solver, functionality, material laws or any of its constituent parts, changes in external dependencies (e.g., software libraries, compilers) or changes in the regulatory framework, requiring re-evaluation of specific credibility factors. Lehman's laws of software evolution postulate that software must continuously evolve to remain useful (Lehman, 1980). This section focuses on the maintenance of the model itself (developer's perspective) rather than the maintenance of specific simulations (user's perspective). It assumes that a model is developed for several uses by different users.

Table 4.6. Good practices for the Maintenance phase of model development on the low-risk and high-risk ends of the risk spectrum.

	Low Risk	**High Risk**
Maintenance	Execution of the maintenance strategy in the model development plan. For each release, report the model's capability to deliver credible results archived together with associated data, documentation, and simulation logs for test cases.	Additionally: continuous recording of incidents, taking corrective, adaptive, perfective, and preventive actions and confirming restored capability according to ISO/IEC/IEEE 12207:2017.

The model development plan should document the maintenance strategy (ISO/IEC/IEEE 12207:2017). Note that medical device regulation (MDR) requires a post-market surveillance plan and periodic safety update reports, which inform about the good practice of maintenance in high-risk contexts (pointing towards continuous monitoring activities).

Release versions should be archived with associated data, documentation, and simulation logs for test cases (NASA-HDBK-7009[21]). Version control is critical for accurate interpretation, repeatability, reproducibility, and debugging of the simulation predictions (Erdemir et al., 2020), and thus for the model's credibility. Ideally, automated tests (see section 4.3.4) are run for each version in continuous integration setups and a standard workflow for releases is defined in continuous deployment setups (NASA-STD-7009A[22]).

As good practice, irrespective of the model risk, new release versions require an additional verification & validation iteration by the developer to guarantee that the released version of the model sustains its credibility for the CoU. In high-risk contexts, continuously monitoring the model's capability to deliver credible results, recording incidents for analysis, taking corrective, adaptive, perfective, and preventive actions, and confirming restored capability (ISO/IEC/IEEE 12207:2017) are likely indicated.

At a certain point, end-of-life decisions will have to be taken. It is important to inform the users in good time about the supported time frame for the model and to clearly communicate which support and training measures are available for users during which phases of the life cycle. Released versions must be archived and preserved in a format that allows execution beyond the supported lifetime. One solution can be software containers (e.g., Docker) that include all dependencies and only rely on an abstract execution layer that will be supported for an extended period. Also, model disposal measures prevent old model versions from returning to the supply chain ISO/IEC/IEEE 12207:2017, unless explicitly required.

4.3. Conclusions

In this Chapter, we outlined an approach to guide model development best practices based on a given CoU. Notice that this Chapter did not address the best practice for the end user of the model directly. Numerous industry standards exist on how to plan, implement, test, and maintain software, as part of medical devices and, thus in critical, regulated environments. As mathematical models for healthcare often take the form of software, the application of an adapted industry standard from software development, for example, ISO/IEC/IEEE 12207:2007 or ICEI/IEC 62304:2006, seems possible. However, full compliance with industry standards is not always required or advisable. We, therefore, suggest using model risk (as defined in Chapter 5) to guide the stringency and level of adherence to industry standards. As a best practice, all models should comply with minimum requirements to anticipate that maximising compliance helps with model / software / platform qualification / certification in regulatory processes.

Life cycle planning reported by a model development plan is suggested as a critical step before implementation. Templates (e.g., SDP for medical device software in regulated environments (The Software Sustainability Institute, 2018) are available and can help to set up high-risk models compliant with current and future regulatory requirements. Requirements must be derived from a

[21] https://standards.nasa.gov/standard/nasa/nasa-hdbk-7009
[22] https://standards.nasa.gov/standard/nasa/nasa-std-7009

detailed analysis of the CoU, the mission and user needs and documented and traced throughout the development.

Of particular importance is the documentation of the model formulation and architecture design decisions, the design itself, and interfaces derived from the requirements as part of a model design document, including a description of intended use cases. Good software development practices should be followed during model implementation and integration, such as version control and the provision of tested code and end-user documentation.

During development and maintenance (as defined in the development plan), integration and systems testing should be performed and reported systematically and automatically. More involved testing paradigms (e.g., unit tests) and continuous monitoring must be envisaged for high-risk environments. Also, testing confirming model credibility for the CoU must be repeated and reported with every model release.

This set of good model development practices provides a general, yet tangible, framework that applies to a wide range of *in silico* models and CoUs spanning different risk levels.

4.4. Essential Good Simulation Practice recommendations

- Establish the CoU(s) of your model, related risks and requirements in a *model development plan* before defining and implementing the model (Table 4.2).
- Identify relevant industry standards for your model (section 4.2).
- When designing the model for your CoU(s), consider relevant domain-specific standards, parameter identifiability and options for verification and validation. Document the decision-making process for the conceptual model and the resulting limitations in the *model design document* (Table 4.3).
- Implement the model software based on established good practices for software engineering and development (Table 4.4) and follow a test-driven development paradigm (Table 4.5).
- Consider the entire model life cycle in the model management plan and secure adequate resources for maintenance (Table 4.6).

5. MODEL CREDIBILITY

Chapter editors:
- Eulalie Courcelles (Novadiscovery)
- Marc Horner (Ansys)
- Payman Afshari (DePuy Synthes, Johnson and Johnson)

Contributors (random order):
- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Steve Kreuzer (Exponent)
- Alexander Kulesza (Novadiscovery)
- Vincenzo Carbone (InSilicoTrials Technologies)
- Cristina Vaghi (InSilicoTrials Technologies)
- Giulia Russo (University of Catania)
- Francesco Pappalardo (University of Catania)
- Raphaëlle Lesage (VPH institute)
- Michael Neidlin (RWTH Aachen University)
- Ghislain Maquer (Zimmer Biomet)
- Philippe Favre (Zimmer Biomet)
- Enrique Morales Orcajo (Ambu Innovation GmbH)
- Liesbet Geris (University of Liège, KU Leuven & VPH Institute)
- Giulia Russo (University of Catania, Italy)
- Francesco Pappalardo (University of Catania, Italy)

5.1. Introduction

The need for a framework to justify that a model has sufficient credibility to be used as a basis for internal or external (typically regulatory) decision-making has been a primary concern with modelling and simulation (M&S) in healthcare. A computational model's credibility is established through verification, validation, uncertainty quantification (VVUQ), and applicability assessment. Verification establishes that a computational model accurately represents the underlying mathematical model and its solution. In contrast, validation establishes whether the mathematical model accurately represents the reality of interest. Uncertainty quantification aids in the identification of potential limitations in the modelling, computational, or experimental processes due to inherent variability (aleatoric uncertainty) or lack of knowledge (epistemic uncertainty). Finally, applicability assesses the relevance of the validation evidence to support using the model for a specific Context of Use (CoU) (Pathmanathan et al., 2017).

Various global organisations have formalised some of these concepts in guidance documents or technical standards for specific use cases. Chapter 1 of this position paper systematically reviews this existing body of knowledge for various computational model types, ranging from QSAR to ABMs to physics-based models. And given the increasing interest in *in silico* methodologies, various global standards bodies (e.g., ICH, ISO) are revising or developing standards in this field.

This chapter outlines concepts related to model credibility assessment in a way that is agnostic to the nature of the computational model type or medical product, extracting the relevant concepts common to the aforementioned standards and guidance documents. To be as inclusive as possible, a level of granularity has been chosen to incorporate most of the existing knowledge, which may result in the grouping or omission of steps that do not exist in every standard or guidance. One example is the EMA's distinction between technical and clinical validation. This was outlined in a letter of support to a request for qualification advice on the use of digital mobility outcomes (DMOs) as monitoring

biomarkers[23], where it was stated that "The technical validation will verify the accuracy of the device and algorithm to measure a range of different DMOs. […] clinical validation will be obtained in an observational multicentre clinical trial" (see also (Viceconti et al., 2020a) for more details). In the follow-up qualification advice, the same authors propose that a DMO is considered clinically validated for a well-defined CoU when one can demonstrate its construct validity, predictive capacity, and ability to detect change (Viceconti et al., 2022). Translating to *in silico* methodologies, both technical and clinical aspects of model predictions must be evaluated during validation. A clinical interpretation of the model validation should also be provided to assess the clinical credibility of the predicted quantities.

This chapter will describe the concepts related to model credibility, supported by illustrative examples and references to standards, guidance, and additional documents that provide further clarification. We also introduce a hierarchical validation approach that distinguishes between a model's physiological, pathological, and treatment layers.

5.2. Model credibility in existing regulatory guidelines

Regulatory agencies have provided some operational guidelines for assessing a predictive model's credibility (see Annex 1 for a complete list of all guidance and standard documents). For example, a 2003 FDA guideline on exposure-response relationships[24] acknowledged that "The issue of model validation is not totally resolved". It recommended (implicitly assuming the models are all data-driven) to separate the training set from the validation set of experimental data. A 2018 EMA guideline on reporting PBPK models[25] recommends validating models against experimental clinical studies of more than 100 patients. It provides instructions on how the comparison between predictions and experiments should be graphed. While this guideline does not explicitly refer to a risk-based credibility assessment, it states: "The acceptance criteria (adequacy of prediction) for the closeness of the comparison of simulated and observed data depends on the regulatory impact". At around the same time, a 2018 FDA guidance on PBPK models[26] requests VVUQ evidence in a generalised sense: "To allow the FDA to evaluate the robustness of the models, the sponsor should clearly present results from the methods used to verify[27] the model, confirm model results, and conduct sensitivity analyses.". However, this guidance also requests that electronic files related to the modelling software and simulations be submitted along with the PBPK study report, to "allow FDA reviewers to duplicate and evaluate the submitted modelling and simulation results and to conduct supplemental analyses when necessary". This may overlook the complexities of reproducing studies involving computational models.

The FDA 2016 guideline on Reporting of Computational Modeling Studies in Medical Device Submissions[28] outlined the importance of providing a complete and accurate summary of computational modelling and simulation evidence that is included in a dossier. This guidance

[23] https://www.ema.europa.eu/en/documents/other/letter-support-mobilise-d-digital-mobility-outcomes-monitoring-biomarkers_en.pdf

[24] https://www.fda.gov/media/71277/download

[25] https://www.ema.europa.eu/documents/scientific-guideline/guideline-reporting-physiologically-based-pharmacokinetic-pbpk-modelling-simulation_en.pdf

[26] https://www.fda.gov/media/101469/download

[27] Please note that the term 'verify' is used in place of validate in the guidance document cited in Footnote 4.

[28] https://www.fda.gov/media/87586/download

referenced the ASME VV-10[29] and ASME VV-20[30] standards but not ASME VV-40 because it was not published then. However, an FDA draft guidance document[31] was published in 2021 that outlines a generalised framework for assessing model credibility that relies heavily upon the ASME VV-40 standard. This guidance proposes ten possible categories of credibility evidence (see table 5.1). It is important to note that categories 1, 4 and 5 are explicitly within the scope of ASME VV-40, while the others may be considered extensions of the ASME VV-40 framework. While the current draft guidance does acknowledge that there are different types of credibility evidence, the issue of different levels of credibility (as proposed in Chapter 2.6) is not considered.

Table 5.1: Ten proposed categories for evidence of credibility. Reprinted from FDA draft guidance "Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions", Dec 2021. The table in the final version of the guidance may differ.

Category	Definition	Definition
1	Code verification results	Results showing that a computational model implemented in software is an accurate implementation of the underlying mathematical model.
2	Model calibration evidence	Comparison of model results with the same data used to calibrate model parameters.
3	General non-CoU evidence	Calculation verification and/or validation evidence gathered for the model under conditions that are broad and not specific to the CoU.
4	Evidence generated using bench-top conditions to support the current CoU	Calculation verification and/or validation evidence using bench-top conditions, that was explicitly planned and generated to support the current CoU.
5	Evidence generated using in vivo conditions to support the current CoU	Same as previous category except using in vivo conditions.
6	Evidence generated using bench-top conditions to support a different CoU	Calculation verification and/or validation evidence using bench-top conditions, that was planned and generated to support a different CoU.
7	Evidence generated using in vivo conditions to support a different CoU	Same as previous category except using in vivo conditions.
8	Population-based evidence	Statistical comparisons of population-level data between model predictions and a clinical data set. (Note: individual-level comparison between model predictions and a clinical dataset falls under Category 5.)
9	Emergent model behaviour	Evidence showing that the model reproduces phenomena that are known to occur in the system at the specified conditions but were not pre-specified or explicitly modelled by the governing equations.
10	Model plausibility	Evidence that supports the validity of the governing equations, model assumptions, and input parameters only.

[29] https://www.asme.org/codes-standards/find-codes-standards/v-v-10-standard-verification-validation-computational-solid-mechanics

[30] https://www.asme.org/codes-standards/find-codes-standards/v-v-20-standard-verification-validation-computational-fluid-dynamics-heat-transfer

[31] https://www.fda.gov/media/154985/download

5.3. A standard framework: ASME VV-40:2018

The American Society of Mechanical Engineers (ASME) Committee on Verification, Validation, and Uncertainty Quantification in Computational Modeling and Simulation has published the ASME VV-10 and VV-20 standards, which outline the processes of verification, validation, and uncertainty quantification for finite element analysis and computational fluid dynamics, respectively (ASME VV-10 2019; ASME VV-20 2009). These standards outline VVUQ best practices but do not provide formalised procedures for steering model validation (and thus the associated model development activities) towards being sufficiently credible for a CoU.

These factors led to the formation of the ASME VVUQ SC 40 subcommittee on Verification, Validation, and Uncertainty Quantification in Computational Modeling of Medical Devices. Through close collaboration between medical device manufacturers, regulatory agencies, and other device industry stakeholders, this subcommittee published the standard "Assessing Credibility of Computational Modeling and Simulation Results through Verification and Validation: Application to Medical Devices" in 2018 (ASME VV-40:2018). This standard introduces a risk-informed credibility assessment framework for physics-based models that applies to various scientific, technical, and regulatory questions. And while the standard is written with a focus on medical devices, the framework is general enough to be translated to a variety of applications, including model-informed drug development and physiologically based pharmacokinetic modelling (Kuemmel et al., 2020; Musuamba et al., 2021; Viceconti et al., 2021b).

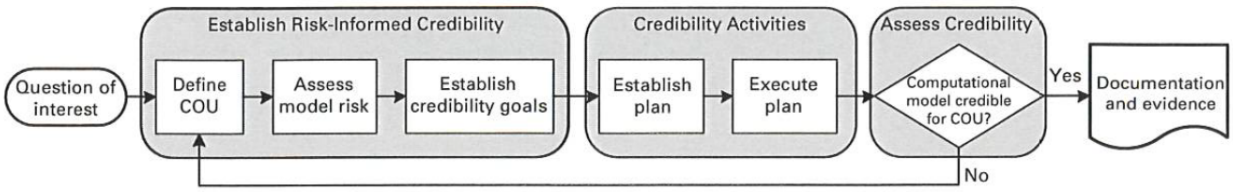


Figure 5.1. Process diagram for the risk-informed credibility assessment framework. ASME VV-40:2018). Reproduced with permission.

The ASME VV-40:2018 risk-informed credibility assessment framework is shown in Figure 5.1. The model credibility assessment begins by stating the question of interest, which describes the specific interrogation, decision, or concern being addressed (at least in part) by the computational model. The next step is to define the CoU, which aims to describe the role and scope of the model fully and exhaustively and how it is going to be used in relation to other forms of evidence, e.g., *in vitro* or *in vivo* data (see (Viceconti et al., 2021a) for examples of CoU), to address the question of interest. The overall model risk is then assessed for the CoU, where risk is a combination of model influence and decision consequence (see Figure 5.2). Model influence is defined as the contribution of the computational model relative to other contributing evidence in deciding, and decision consequence is the consequence (on the patient, for the clinician, business, and/or regulator) if an incorrect decision is made that is based, at least in part, on the model. The overall model risk sets the requirements for model credibility, determining the required degrees of model verification, validation, uncertainty quantification, and applicability such that the model has sufficient credibility for the CoU.

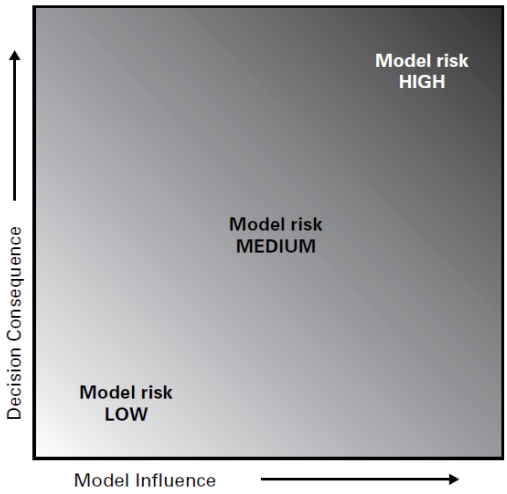


Figure 5.2. Schematic of how model influence and decision consequence determine overall model risk (ASME VV-40:2018). Reproduced with permission.

As an example of how the CoU drives risk-informed credibility, a computational model used for a diagnosis supported by medical imaging and clinical assessment would have lower model risk versus a scenario where the diagnosis relies solely on the computational model. As another example, a model used to define the optimal dosing regimen for a phase 3 clinical trial will have a lower risk if it is complemented with exploratory results from an in vivo phase 2 clinical trial than if used alone. Both scenarios illustrate the impact of model influence on model risk, where the lack of supporting evidence to answer the question of interest means that the model credibility requirements are greater. As an example of the impact of decision consequence, consider a model used to make decisions about a medical device whose adverse outcome could result in severe patient injury or death. In general, this case will be associated with a higher risk than a model used to make decisions regarding a medical device whose adverse outcome would not significantly affect patient safety or health.

Model risk assessment may also be completed with a regulatory impact assessment for certain applications, which describes what evidence would have been provided in the regulatory dossier had it not been for the inclusion of the digital evidence (Musuamba et al., 2020).

5.4. Verification

Verification aims to quantify the part of the predictive error due to the numerical approximations/representations. To effectively separate the three sources of predictive error, the numerical error should be negligible compared to the sum of epistemic and aleatoric errors (see Chapter 3 for details).

There are three possible sources of numerical error in mechanistic models: procedural errors, numerical approximation errors, round-off errors, and numerical discretisation errors. The first two are explored through code verification, while the third is estimated through the calculation verification (Roy and Oberkampf, 2011) (see also ASME VV-10:2019 and ASME VV-20:2009 (R2021)).

5.4.1. *Code verification*

Code verification aims to identify and remove procedural errors in the source code and numerical approximation errors in the solution algorithms. Code verification testing is performed for each computing platform, i.e., hardware configuration and operating system.

Code verification to exclude procedural errors relies on software verification tests such as unit tests, integration tests, and case tests. These tests can be conducted through self-developed or existing automatic software regression suites. In addition, quality control, portability and versioning control should also be considered (see Chapter 3). Moreover, code verification involves developing and implementing a certified software quality assurance (SQA) program to help ensure the integrity of existing code capabilities during development.

To increase rigour in the workflow, ensuring negligible numerical approximation errors, one can use the following code verification methodologies: expert judgement, code-to-code comparisons, discretisation error estimation, convergence studies, and calculating the observed order of accuracy[32]. Apart from expert judgement and code-to-code comparison, each requires comparing code outputs to analytical solutions (or at least mathematical conditions that ensure asymptotic convergence). Traditional engineering problems are one source of analytical solutions, e.g., laminar flow in a pipe or bending of a beam. However, because of their simplicity, these solutions are often limited in their ability to verify the full breadth of the source code. The Method of Manufactured Solutions (MMS) provides a more general source of analytical solutions (Roache, 2019). The Method of Rotated Solutions (MRS) was also recently introduced to expand the scope of traditional engineering problems to provide better code coverage (Horner, 2021). Documented results from verification studies conducted by the software developer may also serve as a source of data to support code verification; however, since numerical accuracy is also hardware-dependent, it is a good practice to repeat those verification tests on the same hardware that will be used to run the models once in use.

Lastly, it is important to note that the scope of the code verification study must include all portions of the simulation platform (e.g., model form, element type, solver) that are accessed as part of validation and model deployment.

5.4.2. *Calculation verification*

Calculation verification (also sometimes referred to as solution verification) aims to estimate the upper bound for the numerical approximation error.

A first important step in calculation verification is to estimate the magnitude of numerical errors caused by the discrete formulation of a mathematical model, e.g., due to iterative errors and discretization errors. The purpose of calculation verification is to analyse the numerical solution's spatial and temporal convergence behaviour by refining the discretisation parameters and convergence tolerances of all iterative schemes and to estimate the numerical errors associated with using a given model.

A sensitivity analysis could also be used to ensure that the calculation does not present a particular combination of input values around which slight variations in the inputs cause significant variations in the outputs (chaoticity). Such occurrence might be due not only to a software bug or insufficiently

[32] https://www.asme.org/codes-standards/find-codes-standards/v-v-10-standard-verification-validation-computational-solid-mechanics

robust solver implementations that would have ideally been caught during code verification but also to ill-conditioning of the numerical problem due to an unfortunate combination of input values.

Finally, calculation verification should ensure that user errors are not corrupting the simulation outputs. No matter how accurate the calculations are, if the result is inaccurately transcribed due for example to a typing error, the predictive model would be totally unreliable.

5.5. Validation

5.5.1. A general definition

As outlined in Chapter 3, validation aims to estimate the prediction error and associated uncertainty of a computational model. An essential part of the validation exercise is the evaluation of the model input(s) and output(s) for the various quantities of interest (QoIs) against a comparator, e.g., the experimental data that are used for validation. Roughly, the comparator should be relevant to the defined CoU and cover a sufficient sample size ("Test samples" in ASME VV-40) as well as the desired range of inputs ("Test conditions" in ASME VV-40). As mentioned in the FDA guidance, acceptable forms of comparators include *in vitro*, *ex vivo*, or *in vivo* test data; these tests may be performed *ex-novo* as part of the validation process (e.g., prospective clinical trial) or based upon historical data (e.g., retrospective clinical trial) or real-world evidence. To evaluate the model capacity to predict the QoIs, this comparator cannot be used during model development or the calibration process.

Additionally, uncertainty quantification is a critical step towards validation of an *in silico* study. This includes estimating the uncertainty associated with the comparator inputs and outputs as well as propagating input uncertainties through the computational model to estimate uncertainty in each QoI.

5.1.1. *Definition and examples*

Mechanistic models rely on four distinct elements: *governing equations* (i.e., the mathematical formulation of the modelled process or phenomenon), *system configuration* (i.e., the device geometry or *in vitro* system), *system properties* (i.e., material properties or physiological parameters) and *system conditions* (i.e., initial, boundary and loading conditions). What is considered as model inputs include initial conditions and parameter values and could extend to:

- part geometry specifications,
- medical imaging settings,
- material framework and ranges,
- boundary conditions for a model of medical device,
- specific patient descriptors such as diet, age, weight, or comorbidities for a drug model,
- numerical parameters used in the four components of the computational model.

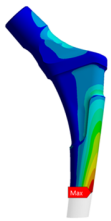
The assessment of the model inputs can be divided into two parts: quantifying sensitivities and uncertainties. The first is concerned with how variations in input parameters propagate through the simulation and their relative impact on the output(s). The sensitivity study results are a rank-order assessment of model input parameters from dominant to negligible impact. On the other hand, quantifying uncertainties addresses how known or assumed uncertainties in the model inputs are

propagated to uncertainties in the model results. The uncertainty analysis provides an error bar (or confidence interval) associated with each model output. In some scenarios, collecting model inputs is the limiting factor in the credibility assessment. However, the same is often true for all evidence collected by designed experiments or observation.

Typically, validation assessment is framed around comparing the model input(s) and output(s) to experimental data - i.e., the comparator - obtained in a set-up that is well-characterised, well-controlled, and relevant to the CoU. This situation corresponds to categories 4 and 6 of credibility evidence outlined in the FDA draft guidance[33] describing sources of model credibility evidence (Table 5.0). The definition of the comparator should include consideration of both the test samples and the test conditions, where each of these can be defined by their quantity, uncertainty, and other descriptors. An assessment of the validation activities should also be used to establish the similarity of model inputs to those of the comparator and the similarity of the outputs and quality of the output comparison.

An example of the various elements of a validation study is provided in Table 5.2.

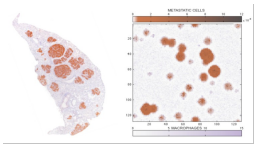
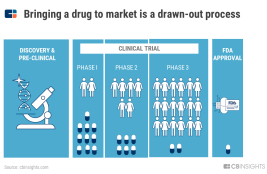
Table 5.2. Example of a validation study for a Finite element analysis (FEA) model predicting the fatigue strength of a hip stem family undergoing ISO 7206-4 standard testing.

Question of Interest	Context of Use (CoU)	Computational model	Comparator	Assessment
Does the proposed hip stem design meet fatigue performance requirements for cyclic loading per ISO 7206-4?	FEA is used to identify the worst-case size hip stem when loaded according to the corresponding ISO 7206-4 standard testing. The FEA model will be used to predict the magnitude and location of the maximum principal stress. The 3 worst-case sizes are referred for bench testing for fatigue performance.	An FEA model of a hip stem family (represented by device geometry and material properties) with system conditions reproducing ISO standard testing.  **Sensitivity** of results to system conditions (such as device position and orientation, load, and potting height/modulus). **Uncertainty** in key stem dimensions, based on manufacturing tolerance, are propagated through the model.	Benchtop testing of the same hip stem family according to the ISO standard testing.  **Test samples:** Statistically relevant number of production parts of 3 different sizes are tested with calibrated equipment. **Test conditions:** A single test condition is considered. Load and displacement are measured without uncertainty quantification in this specific physical test.	Computational model vs Comparator **Input comparison**: All inputs are equivalent (same load, same boundary conditions) **Output comparison**: Ranking of hip stem sizes predicted by the FEA model (peak stress) is compared to ranking in the physical test (fracture load).

[33] FDA draft guidance "Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions", Dec 2021.

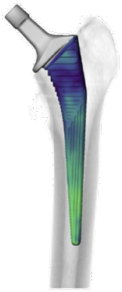
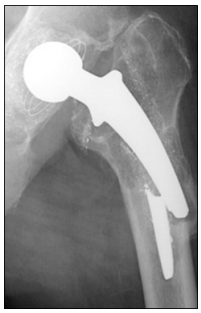
ASME VV-40 provides a framework to demonstrate that a model captures the physics of a medical device by comparison to a well-controlled benchtop test. However, a model used as an *in silico* clinical trial must be shown to reproduce clinical findings. And while the ASME VV-40 standard refers to clinical trials as possible comparators, detailed considerations are not provided. As outlined in the FDA draft guidance[34] on model credibility evidence, a clinical data comparator may be based on *in vivo* tests performed to support a CoU or population-based data from a clinical study (published, retrospective, or prospective) addressing a similar question of interest. And while the clinical comparator does not have to be targeted on the device or drug of interest, it should be reasonably similar to ensure appropriate applicability (see chapter 5.5). Examples of validation for a clinical comparator are provided in Tables 5.3 and 5.4.

Table 5.3. Example of a validation for an agent-based model predicting the proliferation of a known mycobacterium strain in a pulmonary compartment as a function of a vaccine-induced immune response to infection.

Question of Interest	Context of Use (CoU)	Computational model	Comparator	Assessment
What is the vaccine effect size to prevent the active form of tuberculosis disease?	Use an agent-based model of interferon gamma predictions as a response biomarker to select the optimal effective dose of a new therapeutic vaccine against Tuberculosis for which the marketing authorisation is requested. The ABM will be used in the design of phase III trials.	An agent-based model of the human immune system, tuberculosis disease and vaccine mechanism of action.  **Sensitivity** of the results to system conditions (such as bacterial load, immune system profile, and vaccine strategy) is performed.	A randomised double blind phase II clinical trial  **Test samples:** Enrolled patients in the RCT.	Computational model vs Comparator **Input comparison**: All inputs are equivalent. **Output comparison**: Predicted levels of interferon gamma were compared with the comparator output.

[34] FDA draft guidance "Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions,"", Dec 2021.

Table 5.4. Example of validation of an FEA model predicting the fatigue strength of a hip stem during daily living activities.

Question of Interest	Context of Use (CoU)	Computational model	Clinical Comparator	Assessment
Does the proposed hip stem design have sufficient strength to prevent implant fracture in patients?	FEA is used to assess the occurrence of fracture of the hip stem in a virtual patient cohort to enrich clinical data for those configurations which have low (or no) real patient numbers.	An FEA model of a hip stem with system conditions reproducing the expected biomechanical environment and loading conditions during activities of daily living.  **Sensitivity** of the results to the loading conditions and CT-based bone material properties. **Uncertainty** in the cortical bone thickness based on the CT segmentation is propagated through the model.	Stem fracture location and rate in a clinical study of a similar hip stem in a patient cohort.  **Test samples:** Statistically significant number of patients were enrolled following standard practises, covering a wide range of demographics. **Test conditions:** The implant was subjected to a wide range of daily activities, based on clinical scores.	Computational model vs Comparator **Input comparison**: The types of all inputs are similar, but the ranges are not necessarily equivalent (demographics, load) **Output comparison:** Fracture location (visual) and rate were captured in the clinical study and predicted by the FEA models. Fracture rate and location in the virtual cohort were compared to the fracture rate and location in the clinical comparator.

5.5.2. Validation layers for in silico methodologies

To provide rigorous validation of models used to represent preclinical and clinical studies, the ICH E11(R1) guideline on clinical investigation of medicinal products in the paediatric population suggests starting from "pharmacology, physiology and disease considerations.". As such, the following three layers are suggested:

Physiological layer: The model describes the *underlying physiology* of a human or animal system, which could model the treatment at the molecular, cellular, organ, to organ system scale. Associated QoIs are in qualitative or quantitative agreement with an appropriate comparator measured from a healthy system.

Pathological layer: The model describes the *disease processes* of a human or animal system, which could model the treatment at the molecular, cellular, organ, to organ system scale. Associated QoIs are in qualitative or quantitative agreement with an appropriate comparator measured from a pathological system.

Treatment layer: The model describes the *treatment effect* on a physiological or pathological human or animal system (which could model the treatment at the molecular, cellular, organ, to organ system scale). The model could be used to evaluate for a treatment whether the produced QoIs are in qualitative or quantitative agreement with an appropriate comparator.

A strict distinction between these layers is not always possible. For example, the layers may be intertwined in the computational model (e.g., physiological and pathological layers) or even non-existent (e.g., a physiological layer doesn't make sense for the simulation of an *in vitro* experiment).

Most guidelines recommend describing the assessment of model form in detail, defined in VV-40 as "the conceptual and mathematical formulation of the computational model". Where the model is built with these three layers, these need to be described separately and in full detail. But we also recommend, even if this is currently not required by any regulatory guideline or standard, to consider providing results of a validation activity, following the previously described method, for each of these layers, so-called hierarchical validation. Indeed, there is always the theoretical possibility that the errors of one layer hide those of another layer.

5.5.2.1. Uncertainty quantification

The quantification of model uncertainty is achieved by analysing the variation of a specific model output to unknowns occurring in the real-world scenario. Some of them are reducible, but some are inherent. This transforms a deterministic simulation output into a non-deterministic value characterised by a probability and a confidence interval. Performing such analysis is essential within risk-based frameworks for decision-making, as outlined in chapter 5.1. An elegant theoretical framing of the role of uncertainty quantification in decision-making can be found in (Farmer, 2017); reviews of numerical methods for sensitivity analysis used in other industrial sectors can be found in (Cartailler et al., 2014; Schaefer et al., 2020). Some early examples are available for pharmacokinetics models (Farrar et al., 1989), cardiac electrophysiology models (Pathmanathan and Gray, 2014; Pathmanathan et al., 2015; Mirams et al., 2016), models for physiological closed-loop controlled devices for critical care medicine (Parvinian et al., 2019), models of intracranial aneurysms (Sarrami-Foroushani et al., 2016; Berg et al., 2019) and in systems biology models (Villaverde et al., 2022).

Uncertainty quantification is a stepwise process (Roy and Oberkampf, 2011) that begins with identifying all sources of uncertainty, followed by quantifying these uncertainties. Then, uncertainties are propagated through the model to provide the system response, which can be expressed through probabilities under a given confidence interval. A detailed technical explanation and an illustrative example can be found in (Roy and Oberkampf, 2011). The type of uncertainty quantification method (e.g., intrusive, non-intrusive, via surrogate models, etc.) should be chosen appropriately according to the model under investigation (Smith, 2013; Nikishova et al., 2019).

So far, we have implicitly assumed that all model inputs are affected only by a quantification uncertainty, e.g., due to measurement errors. But in some cases, certain inputs are not referred to an individual but rather to a population, and the uncertainty is dominated by inter-subject variability.

Some authors use, in this case, the term "prediction interval", which encapsulates quantification uncertainties and inter-subject variabilities (Tsakalozou et al., 2021)

5.5.2.2. Clinical interpretation of validation results

This section is inspired by EMA's distinction between technical and clinical validation, which is outlined in guidance for "Qualification of novel methodologies for medicine development" [35] and suggested in a letter of support to a request for qualification advice on the use of digital mobility outcomes (DMOs) as monitoring biomarkers. The letter stated, "The technical validation will verify the device's accuracy and algorithm to measure a range of different DMOs. […] clinical validation will be obtained in an observational multicentre clinical trial" (see also (Viceconti et al., 2020a) for more details). While a letter of support is not the most authoritative source, we are unaware of any official source providing such definitions.

The distinction that some regulators make between technical and clinical validation of new methodologies comes from quantitative biomarkers. Technical validation deals with the accuracy with which the quantification is done (i.e., a metrology problem of accuracy and precision estimation); clinical validation deals with the validity of using such measurement as evidence in a specific regulatory decision. Traditionally, the accuracy with which quantitative biomarkers are measured is high, so technical validation is considered necessary but not critical. On the other hand, the relationship between a specific biomarker value and the clinical outcome is usually very complex, so clinical validation is considered the challenging part of assessing a new methodology. The complexity of the relationship between biomarker value and the clinical outcome is also an important reason why clinical validation is usually framed in terms of frequentist statistics. The expectation is that prior knowledge about such a relationship is scarcely informative. Thus, the validity of using the biomarker as a predictor of the clinical outcome is qualified only through an extensive induction, where a very large number of clinical experimental validations are required.

A clinical interpretation of the model validation is an assessment of the clinical credibility of the predicted quantities, i.e., situations where the comparator used in the validation is population-based data collected as part of a clinical trial. This clinical interpretation may be required to satisfy regulatory requirements depending on the CoU. However, there is very little experience and no published guidelines from the regulatory agencies. For the time being, we propose to extract and interpret key results of the validation process similar to the one used to demonstrate the regulatory validity of a conventional biomarker used as an outcome measure.

For conventional biomarkers, which are measured experimentally, an outcome measure is considered valid for a well-defined CoU when one can demonstrate its construct validity, its predictive capacity, and its ability to detect change, where:

- Construct validity is "the extent to which the measure 'behaves' in a way consistent with theoretical hypotheses and represents how well scores on the instrument are indicative of the theoretical construct" (Killewo et al., 2010), page 199. Construct validity is typically demonstrated through the evaluation of simulation results (or model behaviour) regarding what is known (either quantitative data or qualitative knowledge).

[35] https://www.ema.europa.eu/documents/regulatory-procedural-guideline/qualification-novel-methodologies-drug-development-guidance-applicants_en.pdf

- Predictive capacity provides evidence that measures can be used to predict outcomes. It is thus extracted from comparing the model input and output to experimental data obtained in a set-up that is well-characterised, well-controlled, and relevant to the CoU described in the previous section. As written above, experimental data used to demonstrate predictive capability should differ from the data used to develop and calibrate the model.
- The ability to detect change is the most critical aspect, as it relates directly to the decision-making process at the core of the regulatory process. To demonstrate that a prediction can detect change, it is necessary to demonstrate *longitudinal validity*, *minimal important difference*, and *responsiveness*:
    - Longitudinal validity is the extent to which changes in the prediction will correlate with changes in the outcome over time or with changes in measures that are accepted surrogates for the outcome. Whereas predictive capacity is the correlation of the prediction with the outcome at a given time point, longitudinal validity is the correlation of changes in the predictions with changes in the outcome over time. The relationship between the simulated output and the outcome of interest for the regulatory decision should be evaluated as any biomarkers. This relationship may be obvious when the model output is a clinical endpoint or easily supported if the model predicts a validated biomarker (e.g., a validated surrogate endpoint in the example of tuberculosis vaccine efficacy). However, the model would likely need more supportive evidence if the output does not fall into one of these two categories. This question is mostly treated in the definition of the CoU, where the model's use to answer the question of interest is described and justified.
    - The minimal important difference (MID) is the smallest change in the outcome identified as important in the patient's and doctor's opinion. This requires answering the following question: is the model precise enough to detect the MID for the outcome of interest? The answer to this question is extracted from the uncertainty quantification, which gives the prediction confidence interval. The prediction confidence interval cannot include at the same time the MID value and the null value (e.g., absence of difference).
    - Responsiveness to the treatment is the most important attribute for establishing the clinical validity of a predicted biomarker. It can also be described as the model's ability to estimate a clinical benefit. It should be possible to estimate the expected clinical benefit from the estimated predicted change of the simulated result. This last aspect is closely linked to the validation of the treatment layer described above, where the modelling of the treatment impact on the system of interest is evaluated.

Reframing the VVUQ results to this entirely different credibility logic poses several challenges. Many In Silico methodologies can directly predict the primary clinical outcome or at least a QoI that is already accepted as a valid construct for that specific regulatory decision. In this case, construct validity is already ensured. Otherwise, this evidence needs to be generated, using the same approaches used for experimental QoI; for example, by demonstrating convergent and discriminant validity (see for example this systematic review on the topic (Xin and McIntosh, 2017). Predictive capacity and longitudinal validity are two evidences that fit well with the concept of validation according to the ASME VV-40:2018. The concept of minimal important difference is somehow implicit in the VVUQ framework. If the QoI is already accepted in the regulatory practice as a measured value, there is a good chance that a MID value has already been estimated. Again, a MID study is required if the

model predicts a new QoI. The need to demonstrate responsiveness is the one most debated. This typically requires a narrowly defined CoU (specific disease, even a specific range of disease progression, specific class of treatments to be tested) and one or more full randomised interventional clinical trials, possibly conducted by someone independent from the proponents of the in silico methodology.

But all this makes sense only if we consider an L2 validation (see Chapter 3), where the validation expectation is that the model predicts with sufficient accuracy a central property (e.g., the mean) of the distribution of a certain QoI as observed in a well-defined sub-population. Because the implicit assumption is that the accuracy of the predictor may vary as a function of how we choose the validation sub-population, responsiveness assessment requires you to validate with a sub-population that is as close as possible to that you plan to use the in silico methodology (from which the need for a narrow definition of the CoU). But all this would not be valid if the in silico methodology is tested at an L3 level of validity. In this case, each prediction is subject-specific, and the predicted QoI is compared to that measured on the same subject. For such validation, in our opinion, the concept of clinical responsiveness collapses into that Applicability according to the ASME VV-40:2018. Ideally, in an L3 validation study, we want to test the predictive accuracy of the in silico methodology for the widest possible range of patients, severity of the disease, type of treatments, and even across multiple diseases where this is applicable. This is a critical point that will need to be addressed.

5.6. Applicability of the validation activities

Applicability represents the relevance of the validation activities to support using the computational model for the CoU. It includes (i) a systematic review of all validation evidence supporting the use of the model in the CoU, (ii) a precise comparison of the validation context, including both the QOIs and conditions of simulations (e.g., simulated population or experimental conditions and the range of conditions studied) and (iii) a rationale justifying model use despite the potential differences between the validation conditions and the requirements of the CoU. These comparisons are critical since any differences or shortcomings can reduce the overall credibility of the model to answer the question of interest, even in situations where their validation assessment is sufficient. We refer the reader to the framework of Pathmanathan *et al.* (Pathmanathan et al., 2017), which provides step-by-step instructions for determining validation applicability.

In analogy to what is proposed to evaluate the applicability of the analytical validation activities for biomarkers, we recommend describing and assessing the collection/acquisitions, preparation/processing, and storage of the comparator data.[36]

5.7. VVUQ considerations for data-driven models and agent-based models

The logic behind credibility assessment through VVUQ plus applicability analysis assumes implicitly that the model being assessed is mostly knowledge-driven, the knowledge used to build it has resisted extensive falsification attempts, and thus can be considered a scientific truth. Furthermore, the knowledge used to build the model is express in term of mathematical equations. In such a case, the

[36] U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), and Center for Biologics Evaluation and Research (CBER), "Biomarker Qualification: Evidentiary Framework Guidance for Industry and FDA Staff DRAFT GUIDANCE," 2018. https://www.fda.gov/media/119271/download

credibility assessment aims to quantify the prediction error and decompose it into its components (numerical approximation, aleatoric, and epistemic errors). Then the applicability analysis confirms that the prediction error observed in the validation studies represents the prediction error we expect across the range of possible input values.

The extension of this reasoning to data-driven models poses some problems. Here we only give a summary; for an in-depth discussion, please refer to chapter 3. Validation of data-driven models can be performed by calculating the predictive accuracy against one or more annotated datasets (e.g., results of experimental studies where the QoI is measured together with all input values of the model), as far as these datasets were not already used to train the model (test sets). In knowledge-driven models, the epistemic errors are limited to how we implement reliable knowledge in our model; in data-driven models, epistemic errors are not bounded *a priori*. Numerical approximation errors do not exist when there are no equations to solve; hence some verification aspects may not apply (whereas others, such as software quality, remain). Applicability analysis assumes a certain degree of smoothness in how the prediction error varies over the range of possible input values. While for knowledge-driven models, this assumption descends from the properties of the equations that represent the knowledge, such an assumption is not guaranteed in data-driven models. In principle, an artificial neural network model could have a predictive error of 10% of the measured value for a given set of input values, and an error of 100% for another set of inputs, even if those are quite close to the first set. But the bigger difference is related to the risk of *concept drift* that all data-driven model face. Data-driven models make predictions by analysing the correlations between inputs and outputs over a set of experimental measurements. Concept drift means the predictive accuracy of a data-driven model decreases over time. This may happen for several reasons, for example, if the data sample used to train the data-driven model is no longer fully representative of the phenomenon being modelled. While there are techniques to reduce this problem, there is never absolute certainty that concept drift will not occur. This is why there is a growing consensus that the credibility of data-driven must be framed in a Predetermined Change Control Plan, where the predictive accuracy is re-assessed on newly collected data[37].

Agent-based models are a class of predictive models used in biomedicine. These are a generalisation of the concept of cellular automaton first proposed in the 1940's. Most agent-based models are formulated in term rules, through which, at each time step are decided the state transitions of the autonomous agents in the simulation. The key point here is how such rules are defined. If the rules are defined empirically, for example by analysing experimental data, the credibility of that agent-based model should be assessed as for data-driven models, with all the implications mentioned above.

On the contrary, if the rules descend form quantitative knowledge that has resisted extensive falsification attempts, the agent-based model should be considered a knowledge-driven model. However, in this second case, some differences apply, due to the fact that the knowledge that drives the model is not expressed in term of mathematical equations, but in term of rules. This makes the concept of verification more complex to define (see for example (Curreli et al., 2021)).

As this field evolves, more and more sophisticated models will appear. Some problems can be accurately modelled only by combining in a single predictor data-driven modelling and knowledge-driven modelling. The definition of the correct credibility assessment process for such hybrid models is challenging and cannot be generalised. As a rule of thumb, each model should be classified as

[37] https://www.fda.gov/media/145022/download

predominantly data-driven or as predominantly knowledge-driven, and the credibility assessment process should stem from such classification.

5.8. Final credibility

Once the credibility assessment is completed, it must be determined if the model is sufficiently credible for the CoU. Note that the CoU can be modified, and the credibility assessment repeated if the model fails the credibility assessment. Alternatively, the model itself, or the credibility activities, can be revisited and improved to reach the required level of model credibility. A comprehensive summary of the computational model, model results and conclusions must be documented and archived upon conclusion of the modelling project.

5.9. Essential Good Simulation Practice recommendations

- The credibility of in silico methodologies based on predominantly mechanistic models can be effectively demonstrated following the risk-based approach to model verification, validation and uncertainty quantification as detailed in the technical standard ASME VV-40:2018. The credibility of methodologies based on predominantly data-driven models should follow a Predetermined Change Control Plan, where the model's credibility is periodically retested using new test data.
- Where applicable, the validation of predominantly mechanistic models should be done separately for the physiology modelling layer, the disease modelling layer, and for the treatment modelling layer.
- Regulators qualifying in silico methodologies to be used as drug-development tools expect that prior knowledge is generally scarcely informative.
- Regulators currently require that in silico methodologies be used as drug-development tools are qualified following the same regulatory framework used for experimental methodologies. In particular, the technical validation is expected to be separated from the clinical validation. Technical validation deals with the accuracy with which the quantification is done; clinical validation deals with the validity of using such quantity (measured or predicted) as evidence in a specific regulatory decision. Clinical validation requires demonstrating construct validity, predictive capacity, longitudinal validity, minimal important difference, and responsiveness. While some of these concepts map well with the credibility assessment by VVUQ and can be framed simply as special cases of the VVUQ plan, current requirements to demonstrate responsiveness assume an L2 level of validity (population-based validation). This approach is incorrect for in silico methodologies being tested at L3 and should be revised.
- In analogy to what is proposed to evaluate the applicability of the analytical validation activities for biomarkers, we recommend describing and assessing the collection/acquisitions, preparation/processing, and storage of the comparator data used to validate in silico methodologies.

6. POSSIBLE QUALIFICATION PATHWAYS FOR IN SILICO METHODOLOGIES

Chapter editors:

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)

Contributors (random order):

- Cecile F. Rousseau (Voisin Consulting Life Sciences)
- Emmanuelle M. Voisin (Voisin Consulting Life Sciences)
- Alexandre Serigado (Voisin Consulting Life Sciences)

6.1. Introduction

Ultimately, regulatory science is a matter of trust. You need to trust that certain evidence, when obtained with certain methodologies, is sufficient to inform about a new medical product's safety and/or efficacy. Trust is formed based on previous experience but is also informed by the educational background of the experts involved and, in particular, how they decide when a belief can be considered true. And when previous experience is scarce, the educational background drives the decision to trust a new methodology.

Medical device regulators build their regulatory science using an epistemology that is at least in part that of physical sciences. In this context, it is common to expect quantitative experimental results, measurement methodologies mostly free of systematic errors (unbiased), and prior knowledge from fundamental laws of physics and chemistry to be frequently informative. Under these expectations, the inference is mostly Bayesian in that posterior probability is the product of the likelihood probability observed through controlled quantitative experiments and the prior probability that existing knowledge provides. Because the prior knowledge in use has frequently resisted extensive falsification attempts, there is also an expectation that the prior probability and the likelihood are quite similar, which is the theoretical basis of the concept of validation. This opens the door to using in silico methodologies to reduce, refine, and partially replace experimental methodologies.

Drug regulators build their regulatory science using an epistemology proper of natural and social sciences. In this context, it is common to expect experimental results that are qualitative or semi-quantitative. Even when quantitative results are available, there is an expectation that they may be affected by considerable systematic errors caused by selection, information, and confounding biases. There is also the expectation that prior knowledge is scarcely informative due to the complexity and the non-linearity of the phenomena under investigation. Under these expectations, the inference is mostly frequentist. Prior knowledge (and thus in silico methodologies based on it) can, at most, be used to inform the design of experimental studies and to surrogate likelihood probability only when experimental studies are impossible.

As medical products (and the technology to test them) evolve, these expectations need to change. But such change will not happen overnight. The trust in the in silico methodologies will grow as they demonstrate their validity when used as clinical technologies and as clinical research tools in pre-regulatory settings. But in parallel, it is also necessary to break down the cultural walls that separate the regulatory science for medical devices from that for drugs. Scientific advisory panels must become more interdisciplinary, and all expertise should be represented. Targeted re-training programs

are also necessary for the staff of regulatory agencies that inform on the opportunities and risks that innovative technologies pose.

Another possibility discussed in this chapter is to modify the current regulatory qualification pathways for in silico methodologies. This might allow the optimal use of expertise already present within regulatory agencies in providing a thorough and balanced qualification process. In the following sections, we discuss possible alternative pathways to provide elements for reflection to regulatory agencies.

6.2. Pre-certification as Predictive SaMD

Most regulatory authorities nowadays recognise software with a medical purpose as a special class of medical devices called *Software As a Medical Device* (SaMD). The International Medical Device Regulators Forum (IMDRF) defines it as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device." FDA CDRH and the EU CE-marking process both include established regulatory pathways for such technologies. A special case is that of SaMDs with predictive capabilities. Examples of this new class of SaMD are solutions for fractional flow reserve (Zarins et al., 2013), planning software for transcatheter aortic valve replacement (Halim et al., 2021), or software to predict the risk of hip fracture from CT data (Keaveny et al., 2020). A recent FDA draft guideline confirms that even for these solutions, the ASME VV-40: 2018 can be used to assess the credibility of these predictive models.

A first possible regulatory pathway for in silico methodologies could be to require that any evidence supporting the marketing authorisation of a new medical product (whether the medical product is a medical device, a drug, or an ATMP) if obtained *in silico*, should be produced with technologies certified as predictive SaMD. Once the in silico methodology is certified as a predictive SaMD, its qualification as a medical device or drug development tool would be limited to the clinical validation aspects.

The main limitation of this approach is that not all *in silico* methodologies are patient-specific models, and thus their framing into a medical purpose might be impossible. Another potential issue is that some safety requirements that are indispensable for medical purposes might not be necessary when the model is used as an *in silico* methodologies solution; thus, this pathway might be unnecessarily severe for some solutions. On the other hand, it would simplify the regulatory process for solutions intended as SaMD and *in silico* methodologies, as the SaMD certification would cover the technical validation in the qualification process.

6.3. Certification of the technical validity

A more limited version of the SaMD pathway could be a certification of technical validity according to the ASME VV-40:2018 or other similar standards provided by FDA CDRH or EU notified bodies. Once an in silico methodology has such certification, the qualification as a medical device or drug development tool would focus only on clinical validation.

The main limitation of this approach is the need to establish an accreditation process for bodies with the relevant expertise that can produce a credibility certification according to some technical standard.

6.4. Towards an *ad hoc* qualification pathway for *in silico* methodologies

A third possible strategy could be to recognise that the qualification of in silico methodologies, regardless of whether they are used to develop drugs or medical devices, requires a specialised panel. This would imply creating an *ad hoc* process for *in silico* methodologies, which cut through most regulators' current organisation built on the distinction between drugs and devices. The scientific advisory panel would include the same expertise normally found in qualification panels but also experts of in silico methodologies, qualified to evaluate the most technical aspects.

The main limitation of this approach is that an *ad hoc* qualification pathway would need to be created. In the US, such a scenario could be realised through a collaboration between CDER, CBER, and CDRH; the management of such an *ad hoc* qualification pathway could be delegated to one of the three FDA centres or operated under a collaborative model. This would be more complicated in Europe, given that no central authority for medical devices exists.

6.5. Adapting the existing qualification pathways to *in silico* methodologies

The least disruptive approach to the need for a regulatory pathway for in silico methodologies is to embed it into the existing qualification pathways. The FDA provides a qualification pathway for medical device development tools and drug development tools, whereas the EMA provides it only for drug development tools. Qualifying a new methodology is not mandatory, but it is highly recommended, especially for innovative methodologies. Seeking qualification for a method provides an early engagement with the regulatory agencies and will facilitate the integration of this tool into various product development.

Currently, a new methodology is qualified for regulatory use by first requesting qualification advice on the process intended to be used to demonstrate the validity of the new methodology in that CoU. If the authority agrees with the approach, the next step is to conduct the planned validation studies and request a formal qualification opinion. A positive draft qualification opinion is made public for debate if the validity evidence is deemed adequate. If no criticisms emerge from the experts, it is confirmed in its final form. A developer can use that methodology to produce evidence in a marketing authorisation application for a new product without providing additional information on the methodology.

Existing qualification pathways are separated by type of medical product: pathways for drug development tools (e.g., small molecules, biologics, ATMPs, microbiome-derived products), and medical device tools. They currently focus on clinical validation of the methodology rather than on its technical validation. For example, in a recent qualification opinion of EMA on a digital health methodology,[38] the only reference to the technical validity of the new methodology is in a footnote, and the only quantitative requirement is: "The length and velocity of the strides should be accurately measured with an error at 1 sigma (68% confidence interval) under 2.5 %."

One major limitation of this approach is that qualification pathways are not available for all types of products worldwide. Qualification procedures exist both in the EU and in the US. However, it should be noted that while the FDA provides a qualification procedure for new methodologies used to

[38] https://www.ema.europa.eu/documents/scientific-guideline/qualification-opinion-stride-velocity-95th-centile-secondary-endpoint-duchenne-muscular-dystrophy_en.pdf

develop new drugs[39] and new medical devices[40] in the EU, such a pathway is available only for methodologies used to develop new drugs.[41] There are no qualification pathways for medical device development tools in Europe, a major hurdle when advanced, complex, innovative methodologies such as *in silico* methodologies are being proposed. Another issue is the focus of existing scientific advisory panels on the clinical validation aspects. *In silico* methodologies require a thorough credibility assessment, requiring experts to evaluate the dossier properly. Therefore, the use of existing qualification pathways also for *in silico* methodologies would require the extension of the panels to include experts in computational methodologies.

Another issue is that in silico methodologies are sometimes developed to address a specific use case relevant only to that product. In such a case, it would be more convenient to include the evidence of credibility for the in silico methodology in the marketing authorisation dossier rather than undertaking a separate qualification procedure.

6.6. Essential Good Simulation Practice recommendations

- Regulatory agencies should increase the interdisciplinarity of scientific advisory panels and develop targeted re-training programs for their staff on the opportunities and risks that innovative technologies pose.
- Regulatory agencies should explore whether existing qualification pathways need to be adapted to properly include in silico methodologies, or if it would be more convenient to create new qualification pathways for these development methodologies.

39 https://www.fda.gov/drugs/development-approval-process-drugs/drug-development-tool-ddt-qualification-programs

40 https://www.fda.gov/medical-devices/science-and-research-medical-devices/medical-device-development-tools-mddt

41 https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-advice-protocol-assistance/qualification-novel-methodologies-medicine-development-0

7. POSSIBLE HEALTH TECHNOLOGY ASSESSMENT PATHWAYS

Chapter editors:

- Mireia Espallargues Carreras (AQuAS, Agència de Qualitat i Avaluació Sanitàries de Catalunya; RICAPPS, Network for Research on Chronicity, Primary Care, and Health Promotion)
- Xavier Garcia Cuscó (Hospital Universitari Vall d'Hebrón)
- Rossana Alessandrello (AQuAS, Agència de Qualitat i Avaluació Sanitàries de Catalunya)
- Ramon Maspons (AQuAS, Agència de Qualitat i Avaluació Sanitàries de Catalunya)

Contributors (random order):

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Enrique Morales Orcajo (Ambu Innovation GmbH)
- Cecile F. Rousseau (VCLSVoisin Consulting Life Sciences)
- Emmanuelle M. Voisin (VCLSVoisin Consulting Life Sciences)
- Alexandre Serigado (Voisin Consulting Life Sciences)

7.1. Introduction

Two intersections exist between *in silico* methodologies and Health Technology Assessment (HTA). The most obvious is when *in silico* methods are used as predictive Software as Medical Devices (SaMD), e.g., as clinical decision support systems. In such cases, HTA is used, like any other medical technology, to evaluate if its adoption is cost-effective and its clinical appropriateness (criteria under which a certain intervention is properly prescribed to a patient; according to the Italian Medicine Agency (AIFA), appropriateness is defined as "adequacy of the actions adopted to manage a disease, concerning both the patient's needs and the correct use of resources" (Garattini and Padula, 2017)).

A second intersection is when the use of In Silico Trials in the regulatory qualification of a medical product impacts the HTA assessment of that new product. Using In Silico Trials to replace, reduce or refine human experimentation could improve the ability to detect change (which would turn into a more sensitive assessment of differences in efficacy/performance). It could also provide an efficacy/performance assessment closer to real-world effectiveness, as the use of virtual patients may make it easier to explore the efficacy of sub-groups under-represented in clinical trials. In addition, in silico methodologies could reduce or replace Phase 4 trials, produce early estimates of quantities of interest for the HTA assessment of medicinal products, and support the so-called early discourse on HTA. This second perspective is the focus of this chapter.

7.2. Assessing *in silico* methodologies for HTA

In terms of both methodology and application sectors, modelling and simulation is a constantly expanding field. As the models evolve in complexity and increased uptake, it becomes essential to have clarity on the most appropriate tools for the evaluation of in silico methodologies, especially those that can contribute to health technology assessment (HTA).

Developers often make very strong claims with poor reporting and/or a weak verification/validation process of the models. These tools still have a long way to go in terms of implementation and public adoption, as well as rigour in their use, which can be inconsistent and unbalanced at the moment (Musuamba et al., 2021). This chapter aims to provide input on the scientific evaluation of *in silico* methodologies of health interventions (drugs and other technologies) from the HTA point of view and the role that such technologies can play in HTA.

7.3. Introduction to Health Technology Assessment (HTA)

HTA is a multidisciplinary process that uses explicit methods to determine the value of health technology at different points in its lifecycle. The purpose is to inform decision-making to promote an equitable, efficient, high-quality health system[42]. In many countries, it is now common to perform this systematic and multidimensional evaluation of health technologies aimed at informing coverage, reimbursement, or pricing decisions within public healthcare systems.

The process is formal, systematic, and transparent, using state-of-the-art methods to consider the best available evidence. The dimensions of value for a health technology may be assessed by examining the intended and unintended consequences of using a health technology compared to existing alternatives. These dimensions often include clinical effectiveness, safety, costs and economic implications, ethical, social, cultural, and legal issues, organisational and environmental aspects, and wider implications for the patient, relatives, caregivers, and the population. The overall value may vary depending on the perspective taken, stakeholders, and decision context.

HTA can be applied at different points in the lifecycle of health technology, i.e., pre-market, during market approval, post-market, and through the disinvestment of health technology. The approach and methods used in each of these moments will be different and depend on the available evidence (whether primary or secondary data) and the decision to be made about the technology.

Whilst licensing approval is mainly focused on the technical and safety profile of the medical device, HTA bodies have different interests and, therefore, different evidence requirements. Normally, it's aimed at informing policymakers (and decision-makers in general) of the rationale allocation of resources within finite budgets to the funding (or using) of healthcare interventions. For this reason, data required for market access might go beyond those used or developed for licensing, particularly in medical devices, where regulatory requirements have historically been low.

This additional evidence generation could also be worthwhile from the perspective of the manufacturer as with prepaid financing mechanisms for health systems, either through general taxation or private/social insurance, market prospects for medical technologies companies are strongly influenced by third-party payers' coverage. For example, if a CE mark has been granted, this does not imply that the product will be available to patients everywhere in the EU. If the HTA assessment leads to declined public reimbursement in a particular country, the vast majority of patients cannot afford the product in that country.

It is important to mention that "health technology" is a broad concept. The accepted international definition of a health technology is an intervention developed to prevent, diagnose, or treat medical conditions; promote health; provide rehabilitation; or organise healthcare delivery. The intervention

[42] HTA Glossary. International Network of Agencies for Health Technology Assessment (INAHTA), Health Technology Assessment international (HTAi) and other partner organizations. Available at: http://htaglossary.net/HomePage

can be a diagnostic test, device, medicine, vaccine, procedure, program, or system[43]. As we explained in Chapter 2, *in silico* technologies can be used as medical devices, as they are used in the diagnostic, prognostic, or therapeutic process. Otherwise, they can be used to evaluate the safety, efficacy/performance, prescriptive appropriateness, and cost-effectiveness (HTA) of a new medicinal product, whether a medical device or a drug. This chapter will mainly focus on this second use, touching on the first in the final section of future challenges.

Last, it is worth mentioning that modelling and simulation methods are also frequently used to evaluate different types of (implemented) medical interventions, often in the context of HTA. These studies have mainly been used to supplement systematic reviews to increase the usefulness of the evidence summary. Uncertainty about the optimal choice among available interventions for important patient-relevant outcomes may persist even after synthesising the best available evidence. Indeed, decision-makers are increasingly interested in complementing the results of systematic reviews of empirical evidence with information from modelling and simulation studies. That is, integrating empirical evidence on benefits and harms, values (preferences), and/or resource utilisation while accounting for all relevant sources of uncertainty (Dahabreh et al., 2008, 2017). Some of the most frequent applications of this type of modelling and simulation are used:

- to synthesise data from disparate sources (modelling provides mathematical tools for evidence synthesis and the assessment of consistency among data sources),
- to make predictions ("interpolations", forecasts, "extrapolations", prioritisation and planning),
- to support causal explanations and infer the impact of interventions, or
- to inform decision-making (about patient-level care, drug or device licensing, health care policy or the need to conduct additional research (Dahabreh et al., 2017).

Although this specific scenario of modelling and simulation based on the combination of already existing evidence/data could be considered an in silico methodology, it will not be included in this chapter as there are good and updated reviews on that (Dahabreh et al., 2017, 2008; Jalali et al., 2021).

7.4. *In silico* methodologies as a source of evidence

Science generates evidence through observation, deduction, and induction. Simulation, like deduction, starts with specified assumptions regarding a proposed system and generates data suitable for analysis by induction. However, this data does not come from direct observation in the real-world (Stahl, 2008).

These assumptions can be designed according to observed data and predicted as a function of the experimentally observed variability (phenomenological) or by leveraging some pre-existing knowledge about the physics, chemistry, physiology and biology of the phenomenon being modelled (Viceconti et al., 2020b).

In silico methodologies can be a source of evidence when developing or validating a health technology, a pharmaceutical product or a medical device (Model-based medical results or Computational modelling and simulation results). These are predictive computer models that are used to provide evidence in support of the safety and/or efficacy/performance of a medicinal product,

[43] HTA Glossary. International Network of Agencies for Health Technology Assessment (INAHTA), Health Technology Assessment international (HTAi) and other partner organizations. Available at: http://htaglossary.net/HomePage

during its marketing authorisation process. It can also be used during any assessment phase through the technology lifecycle and, thus, become part of the evidence to be used for HTA as well.

Methodologies and tools used to produce regulatory evidence are usually qualified by the regulator or certified according to a specific technical standard such as, for example, the ASME VV-40 for the use of computational modelling to evaluate medical devices[44].

7.4.1. Medical Devices and Interventions

Computational modelling and simulation can help to increase the scientific evidence for evaluating high-risk medical products and interventions, especially when they enable replacing, reducing and refining nonclinical *in vitro* / *ex vivo* experiments, nonclinical animal studies or clinical human studies in case of ethical issues and, time or costs constraints.

It is also particularly significant with the new medical device regulation[45] of the European Commission where scientific evidence used to assess high-risk medical devices must be based on methodologically sound trials, which may be supplemented with alternative evidence sources such as computational modelling and simulation (Olberg et al., 2017).

7.4.2. Pharmaceutical Products

The clearest indication for using simulation methods is when direct experimentation via randomised controlled trials (RCT) is impossible due to cost, time, or ethical constraints. In this regard, RCTs can be considered a form of simulation as it represents and simplifies the system under study. However, computer simulations of these trials typically decrease time and cost, besides overcoming some ethical restrictions of experimentation on humans. These ethical limitations can mainly be found when a question needs exploring (effects of exposure), but conducting the trial would require exposing a vulnerable group to unacceptable risks (Stahl, 2008).

Computational methods aim to complement in vitro and in vivo tests to minimise the need for animal/human testing, reduce the cost and time of toxicity tests, and improve toxicity prediction and safety assessment. *In silico* toxicology encompasses simulation tools for biochemical dynamics and modelling tools for toxicity prediction. They are useful in drug design to determine how drugs should be altered to reduce their toxicity. In turn, this knowledge can be used for the evaluation of pharmaceuticals and to enrich clinical evidence.

For example, there are methods for predicting outcomes based on chemical analogues with known toxicity. On the other hand, researchers also use dose-response or time–response models, which establish relationships between doses or time and the incidence of a defined biological effect (e.g., toxicity or mortality) (Viceconti et al., 2017).

[44] https://www.asme.org/codes-standards/find-codes-standards/v-v-40-assessing-credibility-computational-modeling-verification-validation-application-medical-devices

[45] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745

7.5. *In silico* methodologies: product life cycle and HTA

At the cost of oversimplifying, the development and assessment cycle of any health technology can be reduced into different macro-phases: design/discovery, pre-clinical and clinical assessment, regulatory assessment, market access and post-marketing assessment. Decision maker uncertainty is high in the Discovery and Design phase when new and emerging health technologies have not yet generated any evidence regarding the future value they could bring to the health systems. The more we move through the diffusion curve of technologies, the more evidence is generated and uncertainty reduced. In silico methodologies have the potential to have a role in all steps of the product life cycle (see Table 7.1).

Table 7.1. Potential applications of *in silico* methodologies along the product life cycle and suitable HTA modality.

	Discovery and Design	**Nonclinical and Clinical development**	**Regulatory submission**	**Market Access and Pricing**	**Post-marketing**
Decision maker uncertainty					
HTA modality (see Annex for *Definitions*)	Horizon scanning	Early scientific advice (early dialogues), Pre-Commercial Procurement	Initial HTA, Public Procurement of Innovative Solutions	Mainstream HTA, Coverage and Reimbursement Policy, Value Based Public Procurement	Re-assessment HTA
Potential *in silico* methodologies applications	Streamline target identification & secure proof of concept to broadly explore potential drug combinations and identify those worthy of being progressed into pre- and clinical development Streamline the finding of what new and emerging health technologies have the potentiality to satisfy identified health system unmet needs ➔ Explore efficacy / performance between various targets ➔ Explore efficacy / performance of various	***In silico* methodologies** to be conducted prior to each in vivo clinical trial in order to optimise the design of "real-life" clinical trials ➔ Explore various regimens (dose, number of administration(s) per day, time between administrations, duration of treatment) ➔ Identify subgroups of optimal responders/theranostics biomarkers or generate virtual cohorts with synthetic individuals ➔ Optimise trial design ➔ Optimise health technology trial design to explore its impact in health system quality, sustainability, resiliency, efficiency and equity based on the limited set of trial data	➔ Forecast value to payers by predicting the real-life health technology related benefit and the optimal target use case and population ➔ Benchmark competing innovative health technologies and off the shelf health technologies	➔ Demonstrate value to payers by predicting the real-life technology related benefit and the optimal target population ➔ Transpose Phase 3 trial results into a virtual population representative of specific geographies and context ➔ Demonstrate value to payers by predicting the real-life health technology related benefit and the optimal target use case and population	➔ Reassess value to payers by real world benefit and the optimal target population ➔ Benchmark competing health technologies taking into considerations the market access of new technologies and the achieved effectiveness in the real world

	combinations of targets/treatments ➔ Explore transability from animal models to humans ➔ Explore efficacy / performance of new and emerging health technologies to satisfy identified unmet needs *vs.* doing nothing and keeping current way of doing. ➔ Generate virtual cohorts with synthetic individuals	➔ Explore various regimens and use case of health technology application to limit any kind of bias during the trial once identified the optimal target for which the unmet need is going to be addressed (e.g.: gender, care path determinants, social determinants, access to healthcare data, access to real word data, long term effects, interaction with other interventions, address underrepresented groups, etc.) ➔ Reduce sample size and follow-up duration		➔ Benchmark competing health technologies ➔ To investigate and provide empirical evidence of safety issues	

7.6. Methodologies for *in silico* clinical studies

7.6.1. HTA Health Technology Assessment

Throughout the product life cycle, the industry increasingly relies on computational modelling and simulation to speed development and give further assurance of performance and safety. Still, such use is currently limited (Viceconti et al., 2016). According to the results of a survey of medical device companies in 2014, computational modelling and simulation were more commonly utilised in the early stages of product development or after product commercialisation but rarely to simulate the interaction of the device with a laboratory animal or a patient[46].

When in silico methodologies are used as a source of evidence for health technology development, extending the traditional HTA that informs coverage/reimbursement decisions to early HTA that informs early research, development, and investment decisions (Tummers et al., 2020) could be of great importance, especially for medical devices, where the development process is a costly and uncertain undertaking (Ijzerman and Steuten, 2011). Failed development not only results in a lack of economic return for the company but also in higher costs without healthcare improvements for society. There are multiple reasons for failed device development, but one important factor is the late evaluation of the potential of the device in healthcare practice, usually only after the prototype design is finalised. The aim of the early assessment is to reduce the failure rate at each stage of the development process, while enhancing the efficiency of R&D and of limited resources use, through prioritisation of the innovations most likely to succeed among others. It may also be used to support reimbursement claims by providing quantitative input for developing risk-benefit sharing agreements (Markiewicz et al., 2014). With improved confidence in modelling results and a better-established regulatory framework, the use of in silico evidence as part of the regulatory submission process is becoming more common, but it has not entered the HTA arena yet and evidence from in silico methodologies is seldom used in HTA.

[46] https://avicenna-alliance.com/files/user_upload/Avicenna_Alliance_Position_paper_in_silico_evidence_application_to_Medical_Devices.pdf

7.6.2. *Discovery, design and pre-clinical stages*

The use of *in silico* methodologies in the discovery and design stage can have the potential of streamline target identification; secure proof of concept; identify those drugs/devices worthy of being progressed into pre- and clinical development. Also, streamline the finding of what new and emerging health technologies have the potential to satisfy identified health system unmet needs.

Compared with *in vitro, ex vivo,* and *in vivo* experiments, *in silico* simulations have the advantage of being fast, cheap, safe, easy to implement and free of experimental errors. Consequently, they are becoming increasingly helpful in designing new technologies and strategies.

Simulations and computational models allow the effect of the interaction to be examined not only at the local level, but in the context of the entire pathway in which the target interacts. To include all the features of these complex systems in these pathways, simulation at the biochemical level may be a suitable foundation for simulation. In this sense, different computational models have been proposed to simulate intercellular interaction at the biochemical and physical levels. By means of this type of model, information on the impact of the target on metabolism can be obtained.

Preclinical in silico assays can potentially minimise problems in the translation between experimental and clinical research. Moreover, preclinical data can be a valid source to include in a computational model to gain more insight into the factors that modulate the response in later clinical phases. For this reason, in silico experiments are considered to have a good capacity to make explicit and formalise the underlying mechanisms.

The potential use of *in silico* methodologies can be particularly important in the chemo-prevention and toxicology (Benigni et al., 2020; Valerio, 2009). *In silico* methodologies are used effectively in preclinical studies to optimise dosage administration and predict the overall performance of the optimised schedule (Pappalardo et al., 2019). The number of chemicals marketed for human use is rapidly increasing. For this reason, computational toxicology models have been developed that estimate the event probability of a molecule based on its chemical structure (Quantitative structure-activity relationship or QSAR).

The use of *in silico* experiments to predict toxicological outcomes of drugs and hazard and risk assessment is widespread. Such experiments can determine the priority of molecules for in vivo or in vitro testing. This prioritisation optimises the testing strategy, potentially minimising the need for animal testing (Benigni et al., 2020).

In this regard, Passini et al. have recently developed software which runs *in silico* drug trials in populations of human cardiac models, simulating populations of human action potentials. Designed to predict drug safety and efficacy, the software simulates the effects of drugs on the action potentials of cardiac cells. After conducting variable drug-dose response studies, this software provides statistics of biomarkers of drug action and adverse drug effects, such as arrhythmias. An *in silico* trial of 62 drugs showed that *in silico* simulations predicted clinical risk with 89% accuracy (Passini et al., 2017, 2021). In 2011 the US Food and Drug Administration (FDA) approved the first *in silico* diabetes type 1 model as a possible substitute for pre-clinical animal testing for new control strategies for type 1 diabetes. The European Medicines Agency is considering *in silico* approaches as an alternative to animal testing to protect animal health and the environment.

7.6.3. Clinical development

7.6.3.1. Medical devices

Computational models of the heart based on data obtained from medical imaging of patients have made it possible to use simulations to view different strategies for cardiac rhythm configuration. They have also made it possible to identify the optimal region for localizing cardiac pacing [47]. These models are still in an emerging phase to be considered as medical devices for clinical decision-making. However, this example serves as an illustration of how such models and simulations can be applied in the field of personalized medicine.

Also, as an example of these developments in the field of patient-level simulation, we find the example of blood flow simulation using MRI images and information on blood pressure and blood flow. With the CRIMSON software (Arthurs et al., 2021), 3D models of the blood system could be created. In this way, different surgical strategies could be used to determine a prognosis and to perform an intervention that best preserves blood flow (Ahmed et al., 2021).

The oncNGS pre-commercial procurement[48] aims to develop novel, affordable solutions to provide the best Next Generation Sequencing (NGS) tests for all solid tumours/lymphoma patients. The call for tender[49], launched in December 2021, is challenging the market to address their identified unmet need through the provision of an efficient molecular DNA/RNA profiling of tumour-derived material in liquid biopsies using pan-cancer tumour marker analysis kit including NGS analysis integrated with an in silico decision support system including analytical test interpretation and reporting. The oncNGC PCP contract is structured in three phases:

- Phase 1: Design of the oncNGS solution
- Phase 2: Technical, analytical and clinical performance validation of the oncNGS complete solution prototype at the Supplier's site
- Phase 3: Technical, analytical and clinical performance validation of the oncNGS solution in the clinical samples in Supplier's sites and real clinical settings.

To ensure suppliers keep working on the sustainable dimension of the novel solutions across the three phases, they are required to keep up-to-date *in silico* simulations of their novel panels during both Phase 2 and Phase 3 to demonstrate their solutions are affordable ensuring sufficient and homogeneous coverage of all the targets in agreement with the business case to be applied in routine basis, at each (chemo)therapy cycle to follow clinical response and inspire adaptive therapies.

7.6.3.2. Pharmaceuticals

Phase III clinical trials evaluate a new drug in terms of its clinical value (efficacy and safety), its most appropriate dose and dosage (posology), as well as other aspects such as adherence and tolerability. These *in vivo* studies are expensive and challenging to conduct, as a large sample size is required.

By providing a reliable prediction of the Phase III outcomes based on the data collected during the Phase II clinical trial, *in silico* methodologies may increase confidence in investing at this late stage of the pre-commercial process. High-quality *in silico* methodologies using subject-specific models

[47] Non-invasive simulated electrical and measured mechanical indices predict response to cardiac resynchronization therapy - Research Portal, King's College, London (kcl.ac.uk)

[48] http://oncngs.eu/

[49] https://ted.europa.eu/udl?uri=TED:NOTICE:624705-2021:TEXT:EN:HTML&tabId=1

could be proposed as valid evidence to complement the information from these trials, probably with the requirement to carry out studies to confirm the simulated post-marketing outcomes with real-world data.

By using *in silico* methodologies to predict outcomes for potential phase III trials, it is possible to optimize both the experimental design and the required sample size. As a result, the development cost could be reduced, as well as the time to market (Pappalardo et al., 2019).

7.6.4. Market access and post-marketing assessment

FFR_{CT} software, developed by the US medical firm Heartflow to measure coronary blockages non-invasively from computed tomography scans was the first clinical technology based on subject-specific modelling to get the marketing authorisation from FDA. The software has also received CE marking and regulatory approval in Japan.

What might be relevant from the perspective of decision-makers is the possibility of testing and identifying in advance which patients' subgroups are likely to benefit the most from a novel technology (i.e., enhanced patient population stratification) or to investigate and provide empirical evidence of safety issues that could emerge as a result of the implementation of the technology with consequent streamlined recommendations for a safer and effective indication of use (Ciani et al., 2017).

Another relevant example is the stratification of patients with infectious diseases due to multi-drug resistant (MDR) organisms. Thanks to the provided research and development services contracted through Anti-SUPERBUGS pre-commercial procurement[50], ANTI-SUPERBUGS PCP Buyers' Group aims to:

- Reduce both the costs and the operational impact resulting from infections caused by multi-drug resistant organisms;
- Improve the appropriateness of antimicrobial medicine usage;
- Improve the quality-of-care processes in hospitals;
- Reduce the community and social care impact of MDR infections acquired in hospitals by procuring pre-commercial technologies that will transform current Surveillance and Infections control systems into new comprehensive systems.

The call for tender, launched in November 2019, is challenging the market to address their identified unmet need through the provision of an ASB *in silico* solution comprising a bundle of technologies offering different approaches and outputs at a different level of infection management (as surveillance, environmental safety, first patient screening and patient early diagnosis).

Subsequent public procurement of innovative solutions (PPI), already under preparation, will need to consider that the current COVID-19 pandemic is exacerbating antimicrobial resistance. Data from some EU countries suggest that 6.9% of COVID-19 diagnoses are associated with bacterial infections (3.5% diagnosed concurrently and 14.3% post-COVID-19), with higher prevalence in patients who require intensive critical care (Strathdee et al., 2020). *In silico* methodologies offer the advantage of increasing the cohort, refining clinical validation, and taking into consideration this new potential use case, including intensive care unit (ICU) patients infected by COVID-19 and MDR organisms to be

[50] https://antisuperbugs.eu/

able to demonstrate value to the buyers by predicting the real-life related benefit and the optimal target use case.

7.6.5. *Post-marketing assessment*

In silico studies should also be part of adaptive licensing and reimbursement pathways, where access and coverage are gradually extended as the evidence-based evolves and benefits are demonstrated in clinical research for wider patient populations. This conjunction of interest from regulatory bodies, industry, clinics, academia, and even animal-welfare groups has led to the establishment of networks and initiatives around the world to promote the development, validation, and use of *in silico* medicine technologies.

7.7. Critical assessment of the *in silico* approach and limitations

To present a balanced assessment to consider in future HTA, attention should be drawn to the limitations of current *in silico* simulation tools.

The limitations of in silico simulation techniques should be considered when considering their use in HTA. Primarily, it should be noted that these techniques do not currently allow adequate predictions for all chemicals and outcome variables. Of particular relevance is that there are currently no models for certain systems or components.

The adequacy of the model is of particular interest when evaluating complex systems, such as drugs with multiple mechanisms of action or the interaction of different drugs in poly-medicated patients.

More specifically, in silico experimentation, the limitations derived from the reliability, or transparency, of the data used to design the model on which the simulations will be performed stand out. For example, incorrect training data describing the relationship between dosage and adverse events would amplify these errors in the prediction model.

However, using in silico techniques may add greater uncertainty if it replaces *in vivo* experimentation. This is because assessing the simulation results' external validity is desirable. Recognising the limitations of the technology, there is an increasing interest in combining real-time generated biological data with *in silico* predictions in a rational approach to integrating computational tools with the experimental setting (Jolivette and Ekins, 2007). Using in silico evidence to reduce or refine *in vivo* or *in vitro* experimentation can reduce such uncertainty if reliable and valid models are available.

7.8. How to assess evidence from *in silico* methodologies?

In the context of product development and evaluation, in silico models of increased complexity are often used for similar applications as the 'simpler' pharmacometrics models: trial design optimisation, dose-finding/selection, extrapolation of drug efficacy and safety, etc.

For this reason, considering that the model validation processes described in the previous chapters have been carried out correctly, it is logical that requirements for their acceptability follow the same standards as those already established for models currently included in the regulatory dossier or parallel HTA requests.

To ensure that technology resulting from in silico experiments is properly evaluated, it is important to document these experiments thoroughly. This documentation should allow independent evaluation by HTA bodies in the specific contexts of application of the technology.

When evaluating these experiments, it is important to assess the reliability and relevance of the models used, particularly when the models could pose a risk to patients, involve complex systems, or when there is a considerable distance between the nature of the input (for example, chemical-physical parameters) and the nature or dimension of the output (health symptoms).

Similar to the assessment of *in vitro* evidence, it should be possible to assess the following:[51]

1) The scope of application of the model, the goodness of fit and predictive ability in relation to empirical events in vivo.
2) The validity of the structure and parameters of the model in a biological sense, as well as the degree of modelling of its complexity; moreover, the correspondence of the same or the plausibility of the assumptions.
3) The theoretical basis for the model computations in base.
4) The uncertainty of the model inputs, in the dimensions of natural variability in vivo, reproducibility and reliability.
5) The sensitivity of the model results to the model's uncertainty.
6) The sensitivity of the model results to variations in the input parameters.

To ensure that the model reproduces the results that would be expected *in vivo*, parameters with greater uncertainty and a high or moderate impact on model output should be evaluated. It is also important that the model promotors adequately justify any unexpected results by explaining them based on the model structure, the available data, or the state of understanding of the modelled phenomenon.

To verify that there are no programming or logical errors, the source code of the algorithm and/or operations and processes executed by the model should be made available.

7.9. Challenges for the future

As it was stated at the beginning, *in silico* technologies could also be a health technology itself or part of health technology, that is Digital Patient or Digital Twin technologies. These are predictive computer models that are used as decision-support support systems by a clinician in treating an individual patient, which is in a much more incipient phase of development and assessment. From a regulatory point of view these are considered Software as a Medical Device, but in addition to the specific requirements in terms of Software Quality Assurance such medical devices should be certified for their predictive accuracy, what is called model credibility. Computational modelling and simulation results might eventually be included in regulatory submissions. In that case, the incorporation of these modelling results evidence needs to follow standards of data/evidence generation, analysis, and reporting to enable the regulatory bodies (and HTA agencies) to efficiently perform an adequate assessment of the submitted material.

[51] https://www.oecd.org/chemicalsafety/risk-assessment/guidance-document-on-the-characterisation-validation-and-reporting-of-physiologically-based-kinetic-models-for-regulatory-purposes.pdf

In 2021, the Horizon Europe Framework Horizon program envisages a line of action to provide regulatory agencies and HTA bodies with the necessary tools to exploit the potential of synthetic data[52] for decision making in the field of regulation and health technology assessment. Clearly, one of the challenges of research in this area is to determine the evidence value of this type of information source. Overall, there is a need for rigour and transparency on the one hand in the methods used for *in silico* model development and validation, and on the other hand their wider acceptance as a valuable source of evidence by the scientific community including academic researchers, the pharmaceutical industry, regulatory bodies and HTA/payers (Musuamba et al., 2021). A need exists for documenting the available tools, the manners they are being used, the conditions for their adequate use and the challenges encountered. The current hurdles for the wider acceptability of *in silico* models as a reliable source of evidence for high (HTA) impact applications in drug/medical devices development include:

- lack of common standards and best practice documents commonly accepted by all relevant stakeholders,
- the lack of important digital infrastructure to carry out the *in silico* methodologies (e.g., fast communication networks and high computing power and storage capacity) that could compromise the cost-effectiveness of the resulting health technologies and the coverage, reimbursement or pricing decisions by the public healthcare systems (Leo et al., 2022),
- the protection of individual citizens from the harmful use, also due to security breaches, of their personal data. An approach to solving the challenge surrounding big health data sharing is the generation of synthetic data created from real data by adding statistically similar information,
- biases in algorithm definition and poor training of analysts may pose risks to equity,
- poor communication between stakeholders to that regard,
- the deficit in the skills and knowledge essential to perform HTA based on *in silico* methodologies along the technology life cycle, and
- relatively slower development of regulatory science and HTA as compared to commercial solution developments.

Also, there is currently an unmet need for HTA guidance/best practice documents clearly describing standards for mechanistic *in silico* model development, evaluation and reporting considering the specificities not only in their structure, the data sources for their construction and evaluation but also in the software and algorithms used for their implementation.

Finally, further research is needed to understand the promises of the use of *in silico* methodologies for the development and evaluation of health technologies, to improve their reliability, acceptance, and diffusion and to understand their expected impact on licensing and reimbursement decisions, as well as the full role that can have HTA in the different phases of the application of *in silico* methodologies.

[52] Synthetic data is information that's artificially generated rather than directly captured by real-world events. Typically created using algorithms, synthetic data can be deployed to validate mathematical models and to train machine learning models (https://www.infoq.com/articles/overcoming-privacy-challenges-synthetic-data/)

7.10. Definitions of various HTA modalities

Horizon scanning (Simpson and EuroScan International Network, 2014): this is the systematic identification of new and emerging health technologies that have the potential to impact health, health services, and society; and which might be considered for an HTA. Identification can be:

Proactive: where a range of sources are searched for information on new and emerging health technologies.

Reactive: where systems are in place that allows stakeholders, health professionals, developers and/or consumers to inform the Early Awareness and Alert (EAA) system on new and emerging health technologies

Pre-commercial procurement (PCP)[53,54]: public procurers can drive innovation from the demand side by acting as technologically demanding customers that buy the development and testing of new solutions from several competing suppliers in parallel to compare alternative solution approaches and identify the best value for money solutions that the market can deliver to address their needs. PCP consists of a procurement of Research & Development (R&D) services that involves risk-benefit sharing at market conditions and in which a number of companies develop in competition new solutions for mid-to-long-term public sector needs. The needs are so technologically demanding and in advance of what the market can offer that either no commercially stable solution exists yet, or existing solutions exhibit shortcomings which require new R&D. R&D is split into phases: solution design, prototyping, original development, and validation/testing of a limited set of first products.

Early scientific advice (early dialogues)[55] (Tummers et al., 2020; Ijzerman and Steuten, 2011): is a non-binding scientific advice, before the start of pivotal clinical trials (after feasibility / proof of concept study), in order to improve the quality and appropriateness of the data produced by the developers in view of future HTA assessment / re-assessment. Early HTA is increasingly being used to support health economic evidence development during early stages of clinical research. Such early models can be used to inform research and development about the design and management of new medical technologies to mitigate the risks, perceived by industry and the public sector, associated with market access and reimbursement.

Initial HTA: the early phase HTA helps technology owners or investors make evidence-informed decisions about further investment in the development of medical device and other health technologies, especially with expected public reimbursement or procurement. It attempts to provide appropriate value judgement and assessment of health financing scenarios of innovative technologies before moving ahead with the development process or investing in technology.

Public procurement of innovative solutions (PPI)[56]: PPI happens when the public health systems bodies and providers use their purchasing power to address their identified challenges acting as early adopter of innovative solutions which are not yet available on large scale commercial basis, that are nearly or already in small quantity in the market and don't need new R&D.

---

[53] https://digital-strategy.ec.europa.eu/en/policies/pre-commercial-procurement

[54] {COM(2007) 799 final}, SEC(2007) 1668, COMMISSION STAFF WORKING DOCUMENT accompanying document to the COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Pre-commercial Procurement: Driving innovation to ensure sustainable high quality public services in Europe Example of a possible approach for procuring R&D services applying risk-benefit sharing at market conditions, i.e. pre-commercial procurement, Brussels, 14.12.2007

[55] https://www.eunethta.eu/ja3services/early-dialogues/

[56] https://digital-strategy.ec.europa.eu/en/policies/ppi

Mainstream HTA (Reuzel and Van Der Wilt, 2000): mainstream HTA entails scientific research into the effects and associated costs of health technologies and should support the decision-makers to decide on questions as 'Is this technology better than the technology currently used?', 'How does it compare with alternatives in terms of effectiveness, appropriateness and cost of technologies?' (see chapter 7.3).

Coverage and Reimbursement policy[57]: in decision-making processes regarding the reimbursement of medicines, it needs to be established whether a medicine should be considered eligible for reimbursement. Subsequently, if the medicine is classified as 'reimbursable', it needs to be assessed how much of the price the public payer should (or can) cover. Therefore, setting a price (pricing) and deciding on the level of coverage by public payers (reimbursement) are strongly interlinked. The assessment process usually includes criteria such as efficacy, effectiveness, safety, ease of use, and added therapeutic value, besides cost-effectiveness. In some European countries, the same decision–making process is now used for digital therapeutics[58].

Value-Based Public Procurement[59]: when public health systems bodies and providers adopt off-the-shelf technologies and services and design the relationship with their technology and service providers in agreement to the relationship they have with the payers and to the value the adoption of such off-the-shelf technologies and services bring to the whole healthcare provision chain (from the patients through the payers).

Re-assessment HTA[60]: as the technology matures, changes occur in the technology itself, and there is new evidence available or other factors that can diminish the currency of HTA findings and their utility for health care policies. As such, HTA can be more of an iterative process than a one-time analysis. Coverage and reimbursement policies and subsequent value-based public procurement contracts shall consider the results of HTA reassessments.

7.11. Essential Good Simulation Practice recommendations

In silico methodologies can provide evidence to be used in HTA for:

- demonstrating value to payers by predicting the real-life benefit and the optimal target population for drugs or medical devices
- transposing Phase 3 trial results into a virtual population representative of specific geographies and context
- Benchmarking competing health technologies taking into considerations the market access of new technologies and the achieved effectiveness in the real world.

[57] J. Bouvy, S. Vogler (2013) Update on 2004 Background Paper, BP 8.3 Pricing and Reimbursement Policies, WHO Collaborating Centre for Pharmaceutical Policy and Regulation https://www.who.int/medicines/areas/priority_medicines/BP8_3_pricing.pdf

[58] Driving the digital transformation of Germany's healthcare system for the good of patients - Bundesgesundheitsministerium

[59] Rossana Alessandrello, Ion Arrizabalaga Garde, Uxío Meis Piñeiro, Olman Alonso Elizondo Cordero, Maria Sanchis-Amat, Ramon Maspons (2021). Teoria del canvi, resultats neutrals respecte al tipus de necessitats no satisfetes en l'àmbit de la salut i permeabilitat de les compres públiques d'innovació al valor. Annals de Medicina, vol 104, (2021). Barcelona: Acadèmia de Ciències Mèdiques i de la Salut de Catalunya i de Balears

[60] National Information Center on Health Services Research and Health Care Technology (NICHSR) (2014) HTA 101: Introduction to Health Technology Assessment

8. ETHICAL REVIEW OF *IN SILICO* METHODOLOGIES

Chapter editors:

- Emmanuelle M. Voisin (Voisin Consulting Life Sciences)
- Cecile F. Rousseau (Voisin Consulting Life Sciences)
- Elisabetta Poluzzi (Alma Mater Studiorum – University of Bologna)

Contributors (random order):

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Maria Cristina Jori (Mediolanum Cardio Research)
- Alexandre Serigado (Voisin Consulting Life Sciences))

8.1. Introduction

Before any experimental study can be conducted on humans, the study design must be approved by an independent body responsible for protecting the safety, well-being and rights of the human subjects involved in the experimentation. These bodies are called Independent Ethics Committees in Europe and Institutional Review Boards in the USA; hereinafter, we will use the acronym IEC/IRB to indicate them.

Existing regulatory, legal and ethical frameworks for clinical trials were developed because of well-established medical research practices involving human subjects. Rules were set to protect human research subjects from hazards. By contrast, *in silico* medical research relies on computational resources and data - using patient data to generate and validate computer models, which will be used to predict the necessary evidence.

8.2. Short overview of ethical review in clinical trials

Research involving humans originated in a dark past, where human rights, safety and well-being were disregarded. And that past is not necessarily so remote (see, for example, the Tuskegee Study of Untreated Syphilis in the Negro Male[61]). With the progressive adoption of the Declaration of Helsinki and the establishment of the Good Clinical Practice[62] (GCP), sponsors and Investigators are required to ensure the proper conduct of the clinical trials.

Ethical aspects of any clinical trials are ensured by IEC/IRBs. These entities, which are either local or central, aim to ensure the safety, rights, and well-being of all subjects, whether healthy volunteers or patients, enrolled in any clinical experimentation. Although rules have been originally defined for trials on medicinal products[63], any trial on experimental interventions (*e.g.*, on surgical procedures and on medical devices) must be submitted to IEC/IRBs before starting it. Also, prospective and

[61] https://en.wikipedia.org/wiki/Tuskegee_Syphilis_Study

[62] https://www.ema.europa.eu/en/ich-e6-r2-good-clinical-practice

[63] https://health.ec.europa.eu/medicinal-products/clinical-trials/clinical-trials-regulation-eu-no-5362014_en

retrospective observational studies are submitted to IEC/IRBs to assess risks from additional diagnostic procedures, data protection, and the relevance of the research question.

IEC/IRBs review clinical protocol and the corresponding amendments, the written information on aims, procedures, and rights to be provided to subjects, and the relevant written informed consent forms. They also oversee the enrolment process, including procedures, compensation payments (when appropriate), insurance coverage, the Investigator's qualifications, etc. IEC/IRBs are therefore involved before, during, and after the clinical trial.

8.3. The ethical benefits of *in silico* methodologies

In silico methodologies aim to refine, reduce, and replace experimental studies conducted *in vitro, ex vivo*, or in vivo on animals or humans and provide evidence on medical products' safety, efficacy, and performance.

If we focus on *in silico* methodologies aimed to refine, reduce, and replace human experimentation, several potential ethical benefits can be associated with these new technologies.

8.3.1. Refinement

Refining human experimentation means reducing the risks to which the enrolled subjects are exposed but also increasing the benefit/risk ratio of the experimentation. This means maximising the regulatory utility of the information obtained by exposing the enrolled subjects to such risks. *In silico* methodologies have been proposed to stratify patients better, improving the inclusion and exclusion strategies. This may produce ethical benefits when it helps to identify subjects at higher risk of adverse effects. Where this does not bias the conclusions of the study, such patients can be excluded; alternatively, their identification allows the adoption of measures to mitigate the risk, such as additional monitoring. In some cases, *in silico* methodologies can also directly reduce the risk for enrolled subjects. For example, studies in cardiology that require an invasive fractional flow reserve (FFR) measurement can now be conducted using CT-based virtual FFR models that provide a non-invasive estimate of the FFR for each subject enrolled.

8.3.2. Reduction

When an *in silico* methodology can reduce the number of subjects who need to be enrolled, and thus the number of persons who are exposed to the risks that the study involves, this represents a direct ethical benefit, according to the 6th principle of the Declaration of Helsinki "*In medical research involving human subjects, the well-being of the individual research subject must take precedence over all other interests*". The most obvious examples are *in silico*-augmented clinical trials, where virtual and physical patients are combined (Haddad et al., 2017). Other examples are those cases where the primary outcome is not easily observable, and thus a surrogate biomarker is used to measure response or efficacy. In studies on new drugs to prevent fragility bone fractures caused by osteoporosis, areal bone mineral density is frequently used as a surrogate of the fracture endpoint. A CT-based digital twin can predict the absolute risk of fracture for each patient enrolled; because this predicted quantity has much higher discriminant power, the number of patients enrolled in the clinical studies to achieve statistical power is much smaller (Viceconti and Dall'Ara, 2019).

8.3.3. Replacement

The complete replacement of human experimentation is currently not considered an option. However, there are several cases where human experimentation is impossible and others where a

partial replacement might be an option. Human experimentation is impossible, for example, in assessing the MRI safety of implantable devices (e.g., heating of the device due to high-frequency electromagnetic pulses); here *in silico* methodologies may provide evidence more reliable than animal experiments can provide (Baretta et al., 2020). Scenarios of partial replacement are those where, for example, the digital twin of the subjects enrolled could be used to form a placebo arm in studies where the placebo is considered unethical. Also, in silico methodologies can reduce the numerosity of a clinical required to achieve statistical significance (Haddad et al., 2017). The last scenario where *in silico* methodology may introduce ethical benefits is in studies where, for several reasons, the necessary diversity (of ethnicity, gender, age, physical conditions, etc.) is difficult to achieve with the necessary statistical relevance. *In silico*- augmented clinical trials could be designed not to increase the statistical power of the study but rather to increase it by including tailored virtual patients of such underrepresented sub-groups.

8.4. The ethical review of studies involving *in silico* methodologies

When assessing a medical product involves *in silico* methodologies, are there special attentions that the IEC/IRB need to have in their reviews? To be answered, this general question must be articulated into more specific questions.

What is the role of the IEC/IRB when *in silico* methodologies are used to refine (i.e., to improve rather than to reduce or replace) human studies? We believe that the IEC/IRB is responsible for evaluating if a proper risk analysis has been conducted as part of the *in silico* methodology implementation and if this in silico approach reduces the risks for the subjects enrolled in the clinical trial or helped mitigate the adverse effects in case such risks materialise. In other words, the IEC/IRB needs to evaluate the ethical impact of *in silico* methodologies as they do for any other study methodology. However, this raises an issue of expertise in the current composition of IEC/IRB: such evaluation for *in silico* methodologies may require expertise rarely present in a typical IEC/IRB. In a time where studies involving *in silico* methodologies may still be a rarity, IEC/IRB may circumvent this problem by collecting, in such cases, the opinion of external experts to inform their own decisions. Still, it is reasonable to expect the inclusion of technology experts in IEC/IRB in the long run. Submissions to the IEC/IRB should be extended to include also the technical information necessary to evaluate such *in silico* methodologies.

If *in silico* methodologies are used to reduce the number of subjects enrolled in human studies, we do not see any significant change in how the IEC/IRB operates. In this case, all the concerns are on the reliability of a study's evidence, which concerns the regulatory bodies, not the IEC/IRB. Any means that can reduce the number of subjects enrolled without impacting the statistical relevance of the study should be seen positively from an ethical point of view.

The case where *in silico* methodologies replace human experimentation is the most complex.

The first question is: when a study involves only *in silico* methodologies (for example, in full replacement scenarios), is the IEC/IRB review necessary, considering no human subjects are involved in the study? We believe the answer is no, with one notable exception. IEC/IRBs ensure the safety of human subjects involved in the study; if no human subject is involved, there is no need for the IEC/IRB review. The only exception is when we need to use clinical data to design, inform, or validate the *in silico* methodology. In this case, the IEC/IRB review is required to ensure that the patient's data is treated according to the laws and the ethical principles that regulate these aspects. Frequently the clinical data to be used in the modelling activities are not collected on purpose; this

poses the complex issue of re-using clinical data collected for clinical purposes or for research purposes different from the scope of the current study and whether an additional informed consent of the patients originally involved may be required. Because of its importance, this topic is discussed in greater detail below in a dedicated section.

But as we explained before, the replacement is only partial in some cases. And this frequently occurs when a portion of the study poses ethical problems (e.g., placebo, children, rare diseases). We suggest that such studies should first be subject to a regulatory advice procedure. The regulatory opinion on the appropriateness of the study design[64], including the partial replacement of some human experimentation using *in silico* methodologies, should be acquired by the IEC/IRB, which would focus its evaluation of the ethical implications of the specific implementation of the study design, relying on the regulatory opinion for what matter the reliability of the evidence such study will produce. However, in this case, as in the previous one, the ethical evaluation will be difficult without the involvement of some technology experts. What we wrote before for the refinement scenario is valid also here: while initially, the IEC/IRB may rely on the opinions of external experts, in the long run, it is reasonable to expect the inclusion of technology experts in IEC/IRB.

8.5. Data protection

With real-world data increasing, it is tempting to use them to build and validate computational models. In addition, digital twins in healthcare are informed by the clinical data of individual patients. For such applications, developers must account for data protection laws such as, for example, the European General Data Protection Regulation[65] (GDPR) or the USA Health Insurance Portability and Accountability Act[66] (HIPAA).

An additional complexity for European developers is that the GDPR did acknowledge that the secondary use of clinical data for research purposes could justify some derogation but made no detailed provisions, leaving the member states to define the specific legislation. This has led to a very complex situation, where each country member of the European Union has different legislation. The main problem is not that of privacy (in most cases, the clinical data used in *in silico* methodologies are irreversibly anonymised) but rather that of data ownership. The European GDPR states clearly that the clinical data are owned by the patient, and the clinical institution where the data were generated is allowed to treat these data only for the necessary provision of care. Any secondary use must be explicitly authorised by the patient, the data owner, with informed consent. The point of debate is the granularity of such consent. The orientation of some privacy authorities in EU member states is that consent is given for each research project; thus, if the Investigator plans to reuse the clinical data for another research, he or she needs to collect new informed consent from each patient.

The recent EU Data Governance Act promises to solve this problem. This new EU-wide regulation, which will enter into force in September 2023, provides rules and safeguards to facilitate the re-use of data whenever possible. The main mechanism is that of *data altruism*. *Data altruism* is about individuals and companies giving their consent or permission to make available data that they generate – voluntarily and without reward – to be used in the public interest.

---

[64] As explained in chapter 6, the regulatory pathways for *in silico* methodologies are only partially defined and tend to differ between USA and Europe.

[65] https://gdpr.eu/

[66] https://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996

8.6. Credibility assessment in the IEC/IRB review

In most cases, the IEC/IRB is not called to directly evaluate the evidence of the credibility of the *in silico* methodologies. When the study results are to be used as part of a regulatory submission for marketing authorisation, it is usually expected that before using an *in silico* methodology for a specific context of use, a qualification opinion on such use needs to be obtained by a regulatory agency. In such a case, the qualification opinion should be attached to the IEC/IRB submission. It should be noted that while in the USA, the FDA can provide pathways for the qualification of in silico methodologies for medical devices and drugs development tools, in the EU, such qualification pathway is available only for drug development tools.

However, it could be a good practice to include any evidence of credibility available in the IEC/IRB submission. For example, if the credibility of the *in silico* methodology has been assessed using the ASME VV-40:2018 technical standards, the result summary of this assessment should be included in the submission.

8.7. Essential Good Simulation Practice recommendations

In silico methodologies offer several potential ethical benefits:

- Refining human experimentation means reducing the risks to which the enrolled subjects are exposed but also increasing the benefit/risk ratio of the experimentation, maximising the regulatory utility of the information obtained by exposing the enrolled subjects to such risks.
- When an *in silico* methodology can reduce the number of subjects who need to be enrolled, and thus the number of persons exposed to the study's risks, this represents a direct ethical benefit.
- *In silico* methodologies can provide an ethical alternative where human experimentation is unethical.
- *In silico* methodologies can help in including in clinical studies the necessary diversity (e.g., of ethnicity, gender, age, physical conditions) that, for any reason, might be difficult to achieve experimentally.
- IEC/IRB should evaluate the ethical impact of *in silico* methodologies as they do for any other study methodology. With two special cases, both related to its use to replace human experimentation:
    - For studies where the *in silico* methodologies are used to partially replace human experimentation, the ethical review of the study by the IEC/IRB is necessary. Still, it should be based on the regulatory qualification opinion on the *in silico* methodology.
    - On the contrary, for studies that involve only *in silico* methodologies and no human experimentation, the IEC/IRB review is not necessary, with the notable exception of the ethical management clinical data to design, inform, or validate the *in silico* methodology.
- To properly assess the ethical implications of *in silico* methodologies, IEC/IRB also need technical expertise. Initially, the IEC/IRB may rely on the opinions of external experts. Still, in the long run, it is reasonable to expect the inclusion of technology experts in the IEC/IRB.

9. THE SPONSOR

Chapter editors:

- Maria Cristina Jori (Mediolanum Cardio Research Srl)
- Roberta Bursi (InSilicoTrials Technologies)

Contributors (random order):

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Raphaelle Lesange (VPH institute)
- Axel Loewe (Karlsruhe Institute of Technology)

9.1. Introduction

For the purpose of this document, we define Sponsor as "an individual, company, institution, or organisation that decides to use computer simulations in a preclinical or clinical trial, aimed to a regulatory or decision-making purpose, conducted at any point in a product's lifecycle, both prior to and following marketing authorisation".

According to ICH E6 (R2)[67], the Sponsor is "An individual, company, institution, or organisation which takes responsibility for the initiation, management, and/or financing of a clinical trial". A superimposable definition is given in the standard ISO 14155:2020[68]. Both definitions are confined to the concept of Sponsor in the context of human clinical trials.

As explained in the previous chapters, *in silico* methodologies can refine, reduce, or entirely replace human experimentation. This chapter focuses mainly on studies where in silico methodologies are used to refine or reduce human experimentation; in other words, studies that still involve humans. However, the Sponsor's basic responsibilities are applicable in all contexts, particularly regarding the requirements of implementing a thorough critical-to-quality risk assessment process and to assure the reliability of results. Whatever the aim of the trial and its place in the development path of medical treatment, the Sponsor has an obligation to follow the fundamental principles of Good Clinical Practice (GCP) and/or Good Laboratory Practice (GLP) and/or other GxP, beyond and above the need to follow the present Good Simulation Practice.

In the context of this chapter, when referring to computer simulations or *in silico* trials, we imply the use of models developed and validated according to the requirements covered in Chapters 4 and 5. We also imply that fulfilling all regulatory requirements and guidelines applicable to preclinical and clinical trials related to medicinal products or medical devices is ensured. We will therefore focus on additional requirements to be followed when including computer simulations/*in silico* trials in the development process of new medical treatments.

The Sponsor willing to include *in silico* trials in the frame of the pre-clinical and/or clinical development of a new medical treatment should:

- extensively assess and clearly define the context of use of the *in silico* trial in the development path of its product;
- allocate a project manager and adequate resources;

[67] https://www.ema.europa.eu/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5_en.pdf

[68] https://www.iso.org/standard/71690.html

- identify the computer simulations provider (internal/external);
- draft the trial's protocol;
- analyse regulatory constraints and where necessary seek advice from regulatory authorities;
- ensure continuous oversight of the project;
- critically evaluate the study's outcome and discuss the results with the regulatory authority.

9.2. Relevant expertise

The Sponsor may have internal technical resources/computational specialists or depend on computer simulations vendors/consultants. In any case, the Sponsor should have internal personnel knowledgeable about computer modelling and simulations, at least to the extent needed for adequately assessing technical, regulatory, and logistic constraints. It is recommended that Sponsors with no prior experience in using *in silico* trials put in place a specific implementation plan, including basic training of personnel (e.g., attendance to specific courses, learning of available guidelines and documents, "hands-on" training) or refer to a specialised consultant.

In particular, the Sponsor of an *in silico* clinical trial, whether intended to refine, reduce, or replace human experimentation, should have adequately trained personnel capable of performing the necessary credibility assessment for the *in silico* methodologies, follow available international guidelines and ensure that a quality management system throughout all trial stages is in place.

We do not believe any training certification scheme would be helpful in this case due to the broad range of skills and experience required to use *in silico* methodologies. However, higher education institutions should revise their curricula to include elements of *in silico* medicine in all degrees related to human health. They should also consider more specialised profiles that currently do not exist. The academic experts in *in silico* medicine should collaborate toward defining such curricula.

9.3. Quality management, quality assurance and quality control

The Sponsor of an *in silico* trial should implement a critical-to-quality risk assessment process to ensure:

- the protection of the rights, safety, and well-being of study participants (when these are involved),
- the generation of reliable and meaningful results, and
- the appropriate management of risk factors using a risk-proportionate approach.

9.3.1. Risk identification, evaluation, control, communication, review, and reporting

A basic set of factors relevant to ensuring trial quality should be identified for each study, focusing on critical factors. Examples of possible critical factors are:

- Protocol development: the trial protocol should be scientifically sound and adequately sized, with well-defined and relevant endpoints and statistical methods. Study procedures and conditions for premature study interruption should be detailed. For hybrid studies, measures to protect study participants' rights, safety, and well-being should be defined, in addition to unambiguous

identification of stopping rules for adaptive studies. Studies should also follow the respective good practice documents for the modalities other than in silico (e.g. GCP).

- Selection of the clinical Investigators, as discussed in chapter 9.6.
- Selection of the modeller, as discussed in Chapter 10.
- Trial monitoring/supervision, as discussed under Chapter 10.
- Training of personnel: internal, CRO, and local study staff.
- Data collection and analysis.
- Data interpretation and reporting.

Once identified, the risks should be evaluated regarding the likelihood of occurrence, the extent to which those errors would be detectable and their impact (risk evaluation). Factors identified as critical to quality should be carefully evaluated in advance, and appropriate risk-mitigation activities should be put in place (risk control); in hybrid studies, these should be proportional to the impact of such factors on human subject protection and on the reliability of trial results. Quality management activities and periodic revision and re-assessment of critical factors should be documented. Any change to trial conduct deriving from corrective measures to mitigate critical risks should be documented and reported.

For pilot trials, an external, independent Data Safety Management Board is recommended to set up that periodically reviews data as they accumulate. Studies with adaptive features and/or interim decision points need specific attention during proactive planning, ongoing review of critical quality factors, and risk management.

9.3.2. Standard Operating procedures

The Sponsor should have in place a quality manual and written standard operating procedures (SOPs) to ensure that:

- roles and responsibilities of the personnel (internal/external) are clearly defined and communicated;
- the trial is carried out in compliance with the protocol and applicable regulations. Any deviation from the original plan is recorded, appropriately documented and justified, and its impact on the reliability of the results is properly assessed;
- data generation, data collection, data handling, analysis and reporting are accurately managed to ensure data integrity and reproducibility;
- the process of quality management is defined;
- the process of vendor selection is defined.

9.4. Contract Research Organisation (CRO)

The ICH E6(R2) defines a CRO as “a person or an organisation (commercial, academic, or other) contracted by the sponsor to perform one or more of a sponsor's trial-related duties and functions”. As previously discussed for the definition of Sponsor, in this chapter, the focus will be on the role of

a CRO in the context of human clinical trials. Nonetheless, most of the topics discussed here are general and applicable also to CRO managing pre-clinical trials.

9.4.1. Relevant expertise

The use of CM&S in clinical studies will require a change in the current *status quo* of how CROs operate in drug development projects.

CROs offering services for managing projects that include the adoption of CM&S in any context (like patient-specific models, virtual populations, hybrid trials, in-silico augmented clinical trials, etc.) should have adequate internal staff (highly preferable) or consultants with a good understanding of CM&S, in addition to the expertise in the management of clinical trials. The role of the CRO may or may not include that of developing and running the models. Whenever the CRO also provides computer simulation services, relevant expertise and qualifications, as detailed in Chapter 10, must be ensured.

When the CRO does not have an internal technical department with computational specialists, it might support the Sponsor in identifying the third-party vendor if required by the Sponsor. In all cases, the CRO should have a deep knowledge of applicable regulations, guidelines and best practices related to *in silico* trials and should remain constantly updated as knowledge in the scientific and regulatory fields progresses.

Given the complexity of *in silico* clinical trials, it would be advisable that a specialised professional figure be dedicated to this type of study.

9.4.2. Allocation of roles and responsibilities

The Sponsor may transfer some or all its responsibilities to a CRO but the ultimate responsibility for the quality, and the integrity of the trial remains with the Sponsor. When delegating activities, including in silico activities, the Sponsor's role is to provide the so-called Investigator's Brochure (see Chapter 10) to the mandated Investigator.

The allocation of responsibilities must be in writing, usually in the form of a contract. The Sponsor is also responsible for overseeing the activities performed by the CRO.

Delegated activities may be related to:

- trial design,
- assessment of project feasibility and centres identification,
- model building and development,
- regulatory activities,
- set-up of data collection tools,
- sites initiation and training,
- supervision of the trial conduct (simulations or in human studies),
- site monitoring,
- safety monitoring,

- data handling and data privacy,
- data analysis and reporting,
- maintenance of trial documents.

In addition, when the CRO provides computer simulation services, all responsibilities detailed in Chapter 10 must be fulfilled.

9.5. Adoption of computer simulations in the definition of the global development plan

9.5.1. Pre-clinical development plan

Computer simulation services required to support preclinical studies should be adequately described in a plan, including a description of the *in silico* trial objectives, available knowledge and data, modelling and simulation methodology to be applied, and outcomes evaluation criteria. If the services would be part of application submission to regulatory bodies, Computer simulation activities, including reporting, should be performed according to the recommendations described in Chapters 4 and 5.

CM&S activities should be integral to the sponsor's strategic preclinical development program for the medical product under consideration.

9.5.2. Clinical development plan

Computer simulation services required to support clinical development studies should be performed in line with the recommendations provided in ICH E9 Statistical Principle for Clinical Trials[69] [70]. Specific regulatory guidance documents should be consulted and followed when including model-informed drug development approaches[71].

Modelling activities aiming to analyse the data obtained from a clinical trial should be described in a specific plan, including a description of the objectives, modelling and simulation methodology to be applied, and outcomes evaluation criteria. The *in silico* trial plan should be finalised before the start of the trial. *In silico* trials should be integral to the sponsor's strategic clinical development program for the medical product under consideration.

9.6. Investigator selection

In the context of this chapter, the Investigator is "a person responsible for the conduct of the clinical trial/clinical investigation at a trial site", as defined in the ICH E6(R2) and ISO 14155:2020. Here again, we differentiate the clinical Investigator (i.e., a non-computational specialist) from the modeller, which is discussed in Chapter 10. Although the role and responsibility of the clinical Investigator and the modeller are conceivably different, in the context of hybrid or adaptive clinical trials, the interplay between the two "Investigators" is crucial. There is so far limited experience with

[69] https://www.ema.europa.eu/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf

[70] https://www.ema.europa.eu/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf

[71] Madabushi, R., Seo, P., Zhao, L. et al. Review: Role of Model-Informed Drug Development Approaches in the Lifecycle of Drug Development and Regulatory Decision-Making. Pharm Res 39, 1669–1680 (2022). https://doi.org/10.1007/s11095-022-03288-w

the inclusion of computer simulations in the context of clinical trials in human subjects. As experience is accumulated, such interplay will be formalised.

9.6.1. General requirements

The selection process of a clinical Investigator should consider the context of the use of the *in silico* trial (whether to reduce, refine, or partially replace clinical experiments), the specificities and the complexity of the trial design, and follow a preliminary careful risk evaluation process. In particular, the selection of a clinical Investigator must take into consideration the role and the actual involvement of the Investigator:

- The clinical Investigator is involved in human clinical trials run to validate predictive models,
- The clinical Investigator is involved in a clinical trial simulation (*e.g.*, use of synthetic control arm, virtual populations, digital twins), to inform or to complement the clinical trial,
- The clinical Investigator participates in a hybrid *in silico*/in human trials.

Although a general understanding of modelling and simulation technologies is required in all cases, the level of knowledge in computer simulations the Investigator has should be proportional to the risk: the higher is the risk (which can be quantified with a risks analysis such as the one part of the ASME VV-40:2018 standard), the more qualified should be the Investigator.

Similar considerations apply to Investigator selection in the context of preclinical development.

9.6.2. Investigational centre selection

Based on its role and involvement, the Investigator selection process - in addition to the verification of the requirements established in the ICH E6(R2) and in the ISO 14155:2020 for centre selection - may require the need to perform additional verifications to ensure that the centre has adequate facilities for the *in silico* aspects. It is also important to secure that the Institution and the competent Independent Ethics Committee/Institutional Review Board are well-informed and involved in the process, particularly in the case of complex trial designs.

9.7. Study design, setup, and management

The scope of this section is not to analyse and discuss the different possible designs of an *in silico* trial in the development of a medical product but to provide general guidance and overarching principles.

An in silico trial design should align with the Clinical Development Plan established for that medical product and be preliminarily submitted for advice to regulatory authorities. The regulatory pathway chosen depends on the clinical development plan and the proposed use of the data generated from the *in silico* trial.

The Sponsor should provide an updated Investigator's Brochure detailing all available information related to the medical product, including, in the case, results of performed CM&S.

A study-specific protocol with clearly defined endpoints, a rigorously described methodology, and a proper statistical section must be in place. The rationale and the model's aim (context of use) should

be well described, and the level of the model risk, based on a risk-informed credibility assessment of the computational model. It is recommended that the clinical Investigators are involved in designing the protocol and definitions of the endpoints to ensure that clinical endpoints and engineering outputs are well aligned. The clinical Investigator should also be consulted in preparing the patient's information leaflet and informed consent form, if applicable.

Before the start of the study, ethical and regulatory approvals – as appropriate – are to be obtained. Written agreements among all involved parties (e.g., sponsor, Investigators, institutions, CRO) defining the responsibilities of each party shall be in place.

The general guidelines set in the ICH E6(R2) and in the ISO 14155:2020 should be followed for the study setup, including maintenance of study documents and documentation, the conduct of the study initiation visits and the training of site personnel. The extent of training on computational models for the study site personnel will be customised depending on the specific involvement of the Investigators; in hybrid or adaptive clinical trials, there should be an ongoing interaction between the modeller and the clinical investigator.

The Sponsor should define in a targeted monitoring plan the extent and nature of monitoring appropriate for the study based on risk assessment (see section 9.9).

9.8. Data handling and record keeping

The Sponsor should utilise appropriately qualified (internal or external) individuals to handle and verify the data, conduct the computer simulations analyses, and prepare the trial reports.

For electronic data handling and/or remote electronic trial data systems, the recommendations included in Chapter 5.5 of the ICH E6(R2) should be followed.

The sponsor should retain all sponsor-specific essential documents in conformance with the applicable regulatory requirement(s) of the country(ies) where the product is approved and/or where the sponsor intends to apply for approval(s).

9.9. Compliant GxP Computerised Systems

GxP is an umbrella term that describes regulatory guidelines across the pharmaceutical and medical device industries. The term encompasses a variety of regulatory guidelines such as Good Laboratory Practice (GLPs), Good Clinical Practices (GLPs), Good Manufacturing Practices (GMPs), Good Distribution Practices (GDPs), and Good Storage Practices (GSP).

GxP compliance is establishing and documenting that the specified GxP requirements of a computerised system can be consistently fulfilled. Validation should ensure accuracy, reliability, and consistent intended performance from design until decommissioning of the system or transition to a new system.

Digital systems used for trial purposes should consider the factors critical to their quality in their design and be fit for purpose. To this end, validation of systems, data protection, information technology (IT) security and user management are essential elements to be addressed.

Sponsors should maintain Standard Operating Procedures (SOPs) for using these systems. SOPs should cover system setup, installation, and use. They should further describe system validation and functionality testing, data collection and handling, system maintenance, system security measures,

change control, data backup, recovery, contingency planning, and decommissioning. The responsibilities of the Sponsor, Investigator, and other parties concerning the use of these computerised systems should be clear, and the users should be provided with training in the use of the systems.

Sponsors should further ensure the integrity of the data, including any data that describes the data's context, content, and structure. This is particularly important when changing computerised systems, such as software upgrades or data migration.

The Sponsor may transfer responsibilities of a computerised system to a Technology Service Provider. Still, the ultimate responsibility for the quality and integrity of the computerised system remains with the Sponsor. The allocation of responsibilities must be in writing, usually in the form of a contract. The Sponsor is also responsible for checking that the SOPs of the Technology Service Provider are meeting the Sponsor's quality and integrity standards and overseeing its activities.

9.10. Monitoring procedures

The role of a monitor in the frame of an in silico trial is so far not established. We assume that while a monitor has no role in the verification of the technical aspects of the model, he/she may be involved in ensuring that:

- adequate documentation is produced and maintained during the running of the in silico trial,
- the data used for the models can be tracked to the source,
- the data used for the models are accurate and complete,
- proper informed consent has been obtained from data subjects, where applicable.

Depending on the type of *in silico* trial, these activities should complement standard monitoring activities performed for clinical trials according to current guidelines and regulations to which reference is made.

In all cases, the Sponsor (or delegate) must develop a risk-based monitoring plan based on the risk assessment and tailored to the type and complexity of the study (pre- vs post-market) and its regulatory purpose. In addition to on-site monitoring, centralised monitoring (i.e., a remote evaluation of accumulating data) should be implemented extensively to ensure data quality.

The outcome of all monitoring activities must be documented in the form of reports, which must be timely provided to the Sponsor for review and follow-up.

A special case is when the results of the double-blind clinical experimentation are also to be used to validate the predictive model. In such cases, the clinical data collected during the study have a double use: they inform the safety/efficacy of the new intervention being tested and validate a predictive model. These two activities have different requirements: the analysis to assess safety or efficacy usually takes place once the study is finished, whereas the validation of predictive models may require some of the data (those used as input for the model) to be disclosed to the modeller as soon as they are collected so that the prediction can be made before the validation data are collected (which minimise the risk of bias). This creates a potential issue for the Sponsor, who should be asked to open the labels to the modeller while the trial is still running. A possible solution is this:

- Patients' assignment to study treatments is labelled as groups A and B. The key is disclosed to the modeller only, who is independent of the study team and bound to secrecy;

or

- the input data are identified and stored separated from the rest of the clinical data;
- the modeller is given access to this subset of the clinical data, but no label information;
- the modeller runs the simulations for each patient enrolled twice, once assuming the patient has been treated and one assuming the patient was given the placebo/comparator;
- once the study is completed and the labels are opened, the right simulation is chosen for each patient and compared to the clinically observed values to complete the validation study.

9.11. Audit

One of the critical responsibilities of a Sponsor is to ensure oversight of any clinical trial-related duties and functions, including oversight of the external organisations to which some activities have been delegated (ICH Q10, 21 CFR 211, 21 CFR Part 820.50). The Sponsor should redact an audit plan tailored to the level of risk, focused on critical-to-quality aspects identified in the risk assessment process. Appointed auditors must be independent of the Sponsor and qualified by documented training and experience to conduct audits.

In particular, when computer simulations are outsourced to external vendors, auditors should have the technical expertise to verify critical aspects such as version control for models and software, adherence to standards, and maintenance of adequate documentation.

All findings will be reported to the Sponsor in an audit report to be shared with the audited party. A corrective and preventive action (CAPA) plan should be implemented and followed up for relevant findings.

9.12. Non-compliance

Non-compliance with the protocol, procedures, and regulations can be detected during monitoring or may be a finding from an audit. The Sponsor is responsible for assessing the relevance of the non-compliance and implementing proper corrective actions or terminating the participation of a site/Investigator in the case of serious and /or repeated non-compliance, notifying the regulatory authorities when required by the regulation.

9.13. Premature termination or suspension of a trial

The handling of a premature end of a study involging *in silico* methodologies is quite similar to that used in conventional clinical studies.

The possible reasons for premature termination or suspension of a hybrid/adaptive/in silico-augmented clinical trial should be described in the risk management plan and in the study protocol. If appointed, the independent Data Safety Management Board should be involved in evaluating potentially critical factors. Suppose a decision is made to terminate or suspend a trial, the Investigators and institution. In that case, regulatory authorities and Ethics Committees should be promptly informed and provided the reason(s) for the termination/suspension. The reasons for the premature termination/suspension of an in silico trial not directly involving human subjects should also be documented.

There are fewer reasons for *in silico* trials to terminate early than clinical trials. Nevertheless, this could happen when:

- In an *in silico*-augmented trial, the experimental observations made on the physical subjects enrolled in the study are not consistent with the predictions made for the virtual subjects.
- It becomes clear that the envisioned potential cannot be demonstrated based on an interim analysis.
- The sponsor terminates support and funding based on respective clauses in the agreements.
- The simulation software is not supported anymore by the developers/vendors, and issues or incompatibilities come up that do not allow completing the trial with the existing version. Considering such a scenario, risk should be minimised during model selection/development (see Chapter 4) but cannot be ruled out completely (e.g., bankruptcy).

One of the study arms demonstrates a clear benefit in an interim analysis. Ethically, this does not require trial termination as it could be completed without negative effects after publicising the initial results. Nevertheless, the sponsor could decide to terminate for economic reasons if the intended benefit has been demonstrated already.

Any premature trial termination requires detailed documentation regarding the reasons and circumstances and data acquired and analysed in a dedicated report.

9.14. Trial/study reports

Whether the trial is completed or prematurely terminated, the sponsor should ensure that the trial reports are prepared and, if applicable, provided to the regulatory agency(ies). The sponsor should also ensure that the clinical trial reports are adequately in line with the standards of the ICH E3 Guideline for Structure and Content of Clinical Study Reports[72] and model credibility assessment recommendations provided in Chapter 5.

9.15. Essential Good Simulation Practice recommendations

- The Sponsor of an in silico clinical trial, as well as the CRO that manages it, should have in staff the necessary technical expertise.
- Computer modelling and simulation services required to support clinical development studies should be performed as per the recommendations provided in ICH E9 Statistical Principle for Clinical Trials and be in line with existing regulatory guidelines on the use of CM&S in drug/medical device development plan.
- All computerised systems used in *in silico* clinical studies should be GxP-compliant.

[72] https://www.ema.europa.eu/documents/scientific-guideline/ich-e-3-structure-content-clinical-study-reports-step-5_en.pdf

10. THE INVESTIGATOR: MODELLERS AND ANALYSTS

Chapter editors:

- Raphaelle Lesange (VPH institute)
- Axel Loewe (Karlsruhe Institute of Technology)

Contributors (random order):

- Marco Viceconti (Alma Mater Studiorum – University of Bologna)
- Enrique Morales Orcajo (Ambu Innovation GmbH)

10.1. Roles & responsibilities

In analogy to what is recommended by the ICH guideline on GCP for regular clinical trials, the roles, tasks, and responsibilities of the parties conducting a study involving *in silico* methodologies should be clearly stated and documented appropriately[73,74].

In a clinical study, the Investigator is the person who runs the study. The Investigator may help prepare and carry out the study's protocol (plan), monitor the study's safety, collect and analyse the data, and report the study's results.

When *in silico* methodologies are involved, the term Investigator refers to the person, or in some cases the hosting institution, in charge of carrying out the modelling tasks and generating the *in silico* evidence. Experts who develop predictive models are usually referred to as *modellers*, whereas experts who merely use models developed by others are sometimes called *analysts*. Here we will refer to both roles indistinctly with the term Investigator. Given that the role of the clinical Investigator and that of in silico Investigator may involve different backgrounds, in clinical studies where *in silico* methodologies are involved, the two roles may be separated and assigned to different persons or institutions.

The Investigator may be in charge of performing the simulations and analysis but also of activities described in the model development plan (c.f. Chapter 4) and the credibility-building activities (c.f. Chapter 5). The Investigator's role and responsibilities are defined in relation to the Sponsor and their mutual agreement:

- A documented agreement with the Sponsor should clarify the roles, responsibilities, and frequency of reports at the beginning of the project.
- The Investigator should be aware of and comply with applicable modelling and simulation standards and guidelines, such as the current GSPs.
- The Investigator/institution should have approval of the competent IEC/IRB where required. The Investigator/institution and Sponsor share responsibility for the handling and protection of personal health data, together with the ethics committee.
- The Investigator must follow the model development plan as agreed with the Sponsors. In case of deviation from plan, this should be discussed early on and agreed with the Sponsor in written form. If applicable, new approval and opinion from the ethics committee should be obtained (e.g., when the deviation regards personal health data acquisition, storage, or processing steps).

[73] https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf
[74] https://database.ich.org/sites/default/files/ICH_E6-R3_GCP-Principles_Draft_2021_0419.pdf

- The Investigator is responsible for ensuring that the modelling and simulation activities are carried out with adequate pre-defined hardware and software infrastructures for which the protocol and credibility assessment measures have been designed and approved (c.f. Chapter 4 and Chapter 5).
- Since part of the modelling activities may be delegated to third parties, it is the responsibility of the Investigator or of the Sponsor to record any tasks that have been delegated and the list of the qualified persons they were delegated to. In addition, the Investigator or Sponsor is responsible for adequately informing each third party assisting with the modelling and simulation process about the investigational product (see Investigator's brochure), the modelled system and agreed protocols.

10.2. Investigator's Brochure

Similar to regular clinical trials, the Investigator must be informed by the Sponsor about the medical product under investigation and subjected to the *in silico* trial. This can take place through handing of an investigator's brochure by the Sponsor to the Investigator, like recommended in the ICH Good Clinical Practice[75] and the European Commission Directives 2005/28/EC[76] and 2001/20/EC[77].

The Investigator's brochure summarises the medical product characteristics and compiles existing clinical and non-clinical data (including pre-existing *in silico* data) about the medical products relevant to the study to facilitate understanding the rationale of the *in silico* trial (Döerr et al., 2017).

The information in the investigator's brochure, shall be presented in a concise, simple, objective, balanced and non-promotional form that enables potential investigator to understand it and make an unbiased risk-benefit assessment of the appropriateness of the proposed *in silico* trial.

10.3. Investigator's qualifications

The Investigator needs to be qualified to fulfil their role. The required competencies range from practical skills regarding the use of the simulation software ("know your tools") to the capacity to judge whether the model at hand is suitable for the specific Context of Use (CoU), as detailed below. A lack of general understanding of the physiological processes and the lack of interdisciplinarity in the team are important pitfalls in applied modelling. In particular, the Investigator needs the following qualifications:

- Capacity to judge whether the *in silico* model technique and its boundaries (Intended Use) are compatible with the clinical purpose and objectives (CoU). This assessment requires that the Investigator has access to information on the biomedical context of the study and clinical information about the medical product being modelled. In this context, the Investigator's brochure is of particular importance (see section 10.2).
- Capacity to evaluate the adequacy of the modelling decisions to be taken during the design and the execution of the *in silico* trial and their implications for the intended CoU. This assessment includes biomedical and numerical aspects (for example, time and space resolution, convergence, and stability). When the expertise of the modeller on the pathophysiology of the biomedical

[75] https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf
[76] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32005L0028:EN:NOT
[77] https://health.ec.europa.eu/medicinal-products/clinical-trials/clinical-trials-directive-200120ec_en

process being simulated is not documented, an expert on the specific pathophysiology in question should also be consulted.

- Proficiency in using the M&S software for the *in silico* trial. We refer to Chapter 3 ("Model development") for cases where the software needs to be adapted.
- Capacity to post-process, analyse, and condense the results of the in silico trial, including statistical analysis.
- Capacity to identify relevant ethical aspects related to the *in silico* trial. These can be evaluated by the Investigator or discussed with the institutional ethics committee if required.

Formal training, degrees, and certificates will often be evidence for many of these competencies. However, there is no specific set of degrees of certificates that would be comprehensive enough to cover all aspects and, at the same time, general enough to be applied to all fields of in silico medicine and the wide range of possible Contexts of Use. Considering the wide range of required qualifications, one person is unlikely to fulfil all of them on an expert level. The Investigator needs to ensure that all required competencies are available in the team of experts involved in the study.

The qualifications of the people involved in a simulation study may need to be reported. For instance, in the NASA-hdbk-7009a[78] about CM&S in mission-critical applications, it is requested to "provide an understanding of the education and experience of the people developing and using the M&S" in a dedicated table.

In conclusion, the Investigator needs to convince the relevant stakeholders (Sponsor, regulatory agencies, ethics committee) that the relevant qualifications are available.

10.4. Adequate resources

To execute the in silico trial, or other modelling tasks, agreed on with the Sponsor according to the state of the art, the Investigator needs access to human resources, support, computing resources, and feedback.

In most cases, human resources will be the most expensive part of the *in silico* activities budget. The Investigator and his/her team need to be funded adequately to be able to commit the required time to the execution of the *in silico* experiments and their analysis. The team needs to be formed with persons covering all the required qualifications as detailed in the previous section.

For situations in which the expertise within the team is not sufficient or solutions can be obtained more efficiently with help from the outside, the Investigator should have access to external support. Demand for such support can arise in various fields as evident from the wide range of required qualifications (see section above): technical support from the developers/vendors of the simulation software, support for collecting data, statistical support, support regarding ethical questions or legal and regulatory issues. Resources need to be allocated to pay for such support in case this is not covered by existing agreements. The participation of external experts must be properly documented and tracked throughout the study.

To run the *in silico* experiments the Investigator depends on adequate computing resources. These can range from a personal computer to high-performance computing resources in a dedicated computing centre or in the cloud, depending on the characteristics of the model and the number of

[78] https://standards.nasa.gov/standard/NASA/NASA-HDBK-7009

simulations to be run. The specific requirements need to be discussed and agreed upon with the Sponsor in good time. Remote access needs to be secured according to the state of the art. The choice of the computational platform must also be made keeping in mind the legal and ethical requirements that the treatment of sensitive data imposes.

To ensure that the results of the *in silico* activities will be as valuable as possible, the Investigator should have access to feedback from experts of the biomedical context and/or "users" of the results (e.g., physicians, product managers, regulatory agencies, etc.) during the modelling process unless explicitly designed differently in the study protocol.

10.5. Records and reports

The sequence of steps and decisions made during the modelling process are context-specific and may be subjective, impacting the conclusion and hindering the results' reproducibility (Erdemir et al., 2019). Therefore, concerning a quality approach and/or regulatory evaluation, the M&S tasks and decisions must be documented and reported. Since the Investigator carries out these tasks, here we focus on his/her main responsibilities concerning recording and reporting.

The Investigator must identify, justify, and document every expert-based choice potentially prone to modeller bias (e.g., parameter selection, model structure). All source documents, codes, results, and data should be adequately recorded, maintained, and retained by the Investigator/institution, with the support of the Sponsor, for the duration initially agreed with the Sponsor. It is the responsibility of the Sponsor to agree in advance on an adequate period of time. The tasks that have been delegated should also be subject to recording. In addition, the Investigator must make all records available upon request of the Sponsor or relevant regulatory authorities. As such, the Investigator/institution should ensure the adequate accessibility and legibility of documents and data and support audits.

Regarding reporting activities, the Investigator should provide frequent written progress reports to the Sponsor as defined in the initial agreement. Those reports should document the technical progress and results, as well as potential model deficiencies, limitations and ideas for improvements discovered during the process. Any deviation from the agreed protocol should also be justified and reported by the Investigator when they occur.

Finally, the Investigator must provide the Sponsor and regulatory authorities with a final report summarising the outcome of the *in silico* study after termination. This report should include the actual workflow employed by the Investigator, the generated *in silico* evidence and their analysis concerning the CoU. The Investigator is responsible for the scientific integrity of the reported research and data. The FDA has issued a guidance document providing modellers with a general outline for reporting computational modelling and simulations in medical device submissions[79]. Although detailed content may not entirely apply to all types of in silico models (e.g., for drug approval submission), the general outline is rather generic. It may be considered for guiding the final report of in silico trials. In addition, the EMA provides guidelines to physiologically based pharmacokinetic modellers that describe the expected content of M&S reports for regulatory submissions[80]. Similar guidelines were also released

[79] FDA, Reporting of Computational Modeling Studies in Medical Device Submissions - Guidance for Industry and Food and Drug Administration Staff. (2016). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/reporting-computational-modeling-studies-medical-device-submissions

[80] https://www.ema.europa.eu/en/reporting-physiologically-based-pharmacokinetic-pbpk-modelling-simulation

for reporting population pharmacokinetic analyses[81]. Overall, the content is specific to drug applications, but many recommendations also apply to other modelling applications.

10.6. Safety and security

The major risk for humans is concerning personal data, which need to be handled according to data privacy standards and rules. Given the digital nature of *in silico* trials, there is no direct involvement of human participants from which health-related safety issue could arise during the investigator's modelling activities. For what concerns simulation input measurements and validation activities, any clinical trial that may be necessary to generate data for the model is not the modeller's responsibility and should comply with other relevant guidelines, such as good clinical practices (GCPs). However, the Investigator must consider the following safety and security aspects of *in silico* activities.

Data safety, i.e., protecting data against loss by ensuring safe storage and back-up of the data, must be ensured by the Investigator/institution. This means input (patient) data but also codes, analyses, results, records, and reports. Therefore, appropriate hardware or cloud facilities with backup systems and version control should be available and used by the Investigator (c.f. 9.4). The Investigator should follow the data storage and version tracking strategy as defined in the model development plan (c.f. Chapter 4) with the Sponsor.

Data security is also the responsibility of the Investigator/institution and/or the Sponsor, who should protect personal health data and patient privacy by ensuring adequate use and access restriction to the data. As such, the Investigator based in the European Union must comply with the current General Data Protection Regulation (GDPR)[82] and related directives; most other countries now have similar legislations, although the details may vary considerably. It should be noted that if the country where the data were collected is subject to legislation different from that in force in the country where the data are being treated, the treatment of the data must follow the rules of the country where the data were collected.

Specific attention must be paid to the level of data anonymisation and the possibility of relating some of the data-derived model characteristics (e.g., organ geometry) to the patient's identity. The ethics committee will commonly evaluate the data security strategy and specific measures, which may require a full Data Protection Impact Assessment (DPIA). The Investigator should use patient data according to what was defined in the protocol and approved by the ethics committee.

Finally, any safety issues related to the use of the model within its intended CoU or that emerge as a result of the simulation and/or identified by the Investigator (c.f. Chapter 4) should be detailed in the report to the Sponsor and to the regulatory authorities.

10.7. Essential Good Simulation Practice recommendations

- Role and responsibilities of the investigator are defined in relation to the sponsor and their mutual agreement, which should be documented.
- A record should be kept of eventual third parties contracted to assist in the CM&S activities and they should be adequately informed about the investigational product by the investigator.

[81] https://www.ema.europa.eu/en/reporting-results-population-pharmacokinetic-analyses
[82] https://eur-lex.europa.eu/eli/reg/2016/679/oj

- The investigator needs to ensure and convince stakeholders that all relevant qualifications are available in the team of experts involved in the study.
- The investigator needs access to human resources, support, computing resources, and feedback necessary to accomplish the task as agreed with the sponsor.
- All source documents, codes, results, and data should be adequately version controlled, recorded, maintained, and retained by the investigator/institution, with the sponsor's support, for the duration initially agreed with the sponsor.
- The investigator is responsible for providing regular and final written reports on the conduct of the study and its conclusions by following appropriate reporting guidance.
- The investigator and the sponsor should implement proper data safety and security measures, complying with relevant regulations (GDPR, etc.).

REFERENCES

Ahmed, Y., Tossas-Betancourt, C., van Bakel, P.A.J., Primeaux, J.M., Weadock, W.J., Lu, J.C., Zampi, J.D., Salavitabar, A., Figueroa, C.A., 2021. Interventional Planning for Endovascular Revision of a Lateral Tunnel Fontan: A Patient-Specific Computational Analysis. Front Physiol 12, 718254. https://doi.org/10.3389/fphys.2021.718254

Anzt, H., Bach, F., Druskat, S., Löffler, F., Loewe, A., Renard, B.Y., Seemann, G., Struck, A., Achhammer, E., Aggarwal, P., Appel, F., Bader, M., Brusch, L., Busse, C., Chourdakis, G., Dabrowski, P.W., Ebert, P., Flemisch, B., Friedl, S., Fritzsch, B., Funk, M.D., Gast, V., Goth, F., Grad, J.-N., Hegewald, J., Hermann, S., Hohmann, F., Janosch, S., Kutra, D., Linxweiler, J., Muth, T., Peters-Kottig, W., Rack, F., Raters, F.H.C., Rave, S., Reina, G., Reißig, M., Ropinski, T., Schaarschmidt, J., Seibold, H., Thiele, J.P., Uekermann, B., Unger, S., Weeber, R., 2021. An environment for sustainable research software in Germany and beyond: current state, open challenges, and call for action. https://doi.org/10.12688/f1000research.23224.2

Arthurs, C.J., Khlebnikov, R., Melville, A., Marčan, M., Gomez, A., Dillon-Murphy, D., Cuomo, F., Silva Vieira, M., Schollenberger, J., Lynch, S.R., Tossas-Betancourt, C., Iyer, K., Hopper, S., Livingston, E., Youssefi, P., Noorani, A., Ben Ahmed, S., Nauta, F.J.H., van Bakel, T.M.J., Ahmed, Y., van Bakel, P.A.J., Mynard, J., Di Achille, P., Gharahi, H., Lau, K.D., Filonova, V., Aguirre, M., Nama, N., Xiao, N., Baek, S., Garikipati, K., Sahni, O., Nordsletten, D., Figueroa, C.A., 2021. CRIMSON: An open-source software framework for cardiovascular integrated modelling and simulation. PLoS Comput Biol 17, e1008881. https://doi.org/10.1371/journal.pcbi.1008881

Ashelford, K.E., Weightman, A.J., Fry, J.C., 2002. PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. Nucleic Acids Res 30, 3481–3489. https://doi.org/10.1093/nar/gkf450

Bäker, M., 2018. How to get meaningful and correct results from your finite element model. arXiv:1811.05753.

Baretta, A., Bursi, R., Palazzin, A., Emili, L., 2020. An automated computational tool to study MRI safety of implanted passive cardiovascular medical devices. European Heart Journal 41, ehaa946.3453. https://doi.org/10.1093/ehjci/ehaa946.3453

Benigni, R., Bassan, A., Pavan, M., 2020. In silico models for genotoxicity and drug regulation. Expert Opin Drug Metab Toxicol 16, 651–662. https://doi.org/10.1080/17425255.2020.1785428

Berg, P., Saalfeld, S., Voß, S., Beuing, O., Janiga, G., 2019. A review on the reliability of hemodynamic modeling in intracranial aneurysms: why computational fluid dynamics alone cannot solve the equation. Neurosurg Focus 47, E15. https://doi.org/10.3171/2019.4.FOCUS19181

Byon, W., Smith, M.K., Chan, P., Tortorici, M.A., Riley, S., Dai, H., Dong, J., Ruiz-Garcia, A., Sweeney, K., Cronenberger, C., 2013. Establishing best practices and guidance in population modeling: an experience with an internal population pharmacokinetic analysis guidance. CPT Pharmacometrics Syst Pharmacol 2, e51. https://doi.org/10.1038/psp.2013.26

Cartailler, T., Guaus, A., Janon, A., Monod, H., Prieur, C., Saint-Geours, N., 2014. Sensitivity analysis and uncertainty quantification for environmental models. ESAIM: Proc. 44, 300–321. https://doi.org/10.1051/proc/201444019

Ciani, O., Federici, C., Pecchia, L., 2017. The evaluation of medical devices: are we getting closer to solve the puzzle? A review of recent trends, in: Proceedings of EMBEC & NBC 2017. Presented at the Joint Conference of the European Medical and Biological Engineering Conference (EMBEC) and the Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC), Tampere, Finland.

Cucurull-Sanchez, L., Chappell, M.J., Chelliah, V., Amy Cheung, S.Y., Derks, G., Penney, M., Phipps, A., Malik-Sheriff, R.S., Timmis, J., Tindall, M.J., van der Graaf, P.H., Vicini, P., Yates, J.W.T., 2019. Best Practices to Maximize the Use and Reuse of Quantitative and Systems Pharmacology Models: Recommendations From the United Kingdom Quantitative and Systems Pharmacology Network. CPT: Pharmacometrics & Systems Pharmacology 8, 259–272. https://doi.org/10.1002/psp4.12381

Curreli, C., Pappalardo, F., Russo, G., Pennisi, M., Kiagias, D., Juarez, M., Viceconti, M., 2021. Verification of an agent-based disease model of human Mycobacterium tuberculosis infection. Int J Numer Method Biomed Eng 37, e3470. https://doi.org/10.1002/cnm.3470

Dahabreh, I.J., Chan, J.A., Earley, A., Moorthy, D., Avendano, E.E., Trikalinos, T.A., Balk, E.M., Wong, J.B., 2017. Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment, AHRQ Methods for Effective Health Care. Agency for Healthcare Research and Quality (US), Rockville (MD).

Dahabreh, I.J., Trikalinos, T.A., Balk, E.M., Wong, J.B., 2008. Guidance for the Conduct and Reporting of Modeling and Simulation Studies in the Context of Health Technology Assessment, in: Methods Guide for Effectiveness and Comparative Effectiveness Reviews, AHRQ Methods for Effective Health Care. Agency for Healthcare Research and Quality (US), Rockville (MD).

Döerr, B., Whitman, S., Walker, S., 2017. Writing for medical devices compared to pharmaceuticals: An introduction. MEW 26, 8–13.

Erdemir, A., Besier, T.F., Halloran, J.P., Imhauser, C.W., Laz, P.J., Morrison, T.M., Shelburne, K.B., 2019. Deciphering the “Art” in Modeling and Simulation of the Knee Joint: Overall Strategy. J Biomech Eng 141, 0710021–07100210. https://doi.org/10.1115/1.4043346

Erdemir, A., Mulugeta, L., Ku, J.P., Drach, A., Horner, M., Morrison, T.M., Peng, G.C.Y., Vadigepalli, R., Lytton, W.W., Myers, J.G., 2020. Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective. Journal of Translational Medicine 18, 369. https://doi.org/10.1186/s12967-020-02540-4

Farmer, C.L., 2017. Uncertainty quantification and optimal decisions. Proc Math Phys Eng Sci 473, 20170115. https://doi.org/10.1098/rspa.2017.0115

Farrar, D., Allen, B., Crump, K., Shipp, A., 1989. Evaluation of uncertainty in input parameters to pharmacokinetic models and the resulting uncertainty in output. Toxicol Lett 49, 371–385. https://doi.org/10.1016/0378-4274(89)90044-1

Frigg, R., Hartmann, S., 2020. Models in Science [WWW Document]. The Stanford Encyclopedia of Philosophy (Spring 2020 Edition). URL https://plato.stanford.edu/archives/spr2020/entries/models-science/

Galappaththige, S., Gray, R.A., Costa, C.M., Niederer, S., Pathmanathan, P., 2022. Credibility assessment of patient-specific computational modeling using patient-specific cardiac modeling as an exemplar. PLOS Computational Biology 18, e1010541. https://doi.org/10.1371/journal.pcbi.1010541

Garattini, L., Padula, A., 2017. 'Appropriateness' in Italy: A 'Magic Word' in Pharmaceuticals? Appl Health Econ Health Policy 15, 1–3. https://doi.org/10.1007/s40258-016-0240-7

Haddad, T., Himes, A., Thompson, L., Irony, T., Nair, R., 2017. Incorporation of stochastic engineering models as prior information in Bayesian medical device trials. Journal of biopharmaceutical statistics 27, 1089–1103. https://doi.org/10.1080/10543406.2017.1300907

Halim, J., Brouwer, J., Lycke, M., Swaans, M.J., Van der Heyden, J., 2021. Transcatheter aortic valve replacement: impact of pre-procedural FEops HEARTguide assessment on device size selection in borderline annulus size cases. Neth Heart J 29, 654–661. https://doi.org/10.1007/s12471-021-01620-4

Horner, M., 2021. The Method of Rotated Solutions: A Highly Efficient Procedure for Code Verification. Journal of Verification, Validation and Uncertainty Quantification 6. https://doi.org/10.1115/1.4049322

Ijzerman, M.J., Steuten, L.M.G., 2011. Early assessment of medical technologies to inform product development and market access: a review of methods and applications. Appl Health Econ Health Policy 9, 331–347. https://doi.org/10.2165/11593380-000000000-00000

Jalali, M.S., DiGennaro, C., Guitar, A., Lew, K., Rahmandad, H., 2021. Evolution and Reproducibility of Simulation Modeling in Epidemiology and Health Policy over Half a Century. Epidemiol Rev mxab006. https://doi.org/10.1093/epirev/mxab006

Jean, D., Naik, K., Milligan, L., Hall, S., Mei Huang, S., Isoherranen, N., Kuemmel, C., Seo, P., Tegenge, M.A., Wang, Y., Yang, Y., Zhang, X., Zhao, L., Zhao, P., Benjamin, J., Bergman, K., Grillo, J., Madabushi, R., Wu, F., Zhu, H., Zineh, I., 2021. Development of best practices in physiologically based pharmacokinetic modeling to support clinical pharmacology regulatory decision-making—A workshop summary. CPT: Pharmacometrics & Systems Pharmacology 10, 1271–1275. https://doi.org/10.1002/psp4.12706

Jolivette, L.J., Ekins, S., 2007. Methods for predicting human drug metabolism. Adv Clin Chem 43, 131–176. https://doi.org/10.1016/s0065-2423(06)43005-5

Ke, A., Barter, Z., Rowland-Yeo, K., Almond, L., 2016. Towards a Best Practice Approach in PBPK Modeling: Case Example of Developing a Unified Efavirenz Model Accounting for Induction of CYPs 3A4 and 2B6. CPT Pharmacometrics Syst Pharmacol 5, 367–376. https://doi.org/10.1002/psp4.12088

Keaveny, T.M., Clarke, B.L., Cosman, F., Orwoll, E.S., Siris, E.S., Khosla, S., Bouxsein, M.L., 2020. Biomechanical Computed Tomography analysis (BCT) for clinical assessment of osteoporosis. Osteoporos Int 31, 1025–1048. https://doi.org/10.1007/s00198-020-05384-2

Killewo, J., Heggenhougen, K., Quah, S.R., 2010. Epidemiology and Demography in Public Health. Academic Press.

Kuemmel, C., Yang, Y., Zhang, X., Florian, J., Zhu, H., Tegenge, M., Huang, S.-M., Wang, Y., Morrison, T., Zineh, I., 2020. Consideration of a Credibility Assessment Framework in Model-Informed Drug Development: Potential Application to Physiologically-Based Pharmacokinetic Modeling and Simulation. CPT: Pharmacometrics & Systems Pharmacology 9, 21–28. https://doi.org/10.1002/psp4.12479

Lehman, M.M., 1980. Programs, life cycles, and laws of software evolution. Proceedings of the IEEE 68, 1060–1076. https://doi.org/10.1109/PROC.1980.11805

Leo, C.G., Tumolo, M.R., Sabina, S., Colella, R., Recchia, V., Ponzini, G., Fotiadis, D.I., Bodini, A., Mincarone, P., 2022. Health Technology Assessment for In Silico Medicine: Social, Ethical and Legal Aspects. International Journal of Environmental Research and Public Health 19, 1510. https://doi.org/10.3390/ijerph19031510

Markiewicz, K., van Til, J.A., IJzerman, M.J., 2014. Medical devices early assessment methods: systematic literature review. Int J Technol Assess Health Care 30, 137–146. https://doi.org/10.1017/S0266462314000026

Massoud, T.F., Hademenos, G.J., Young, W.L., Gao, E., Pile-Spellman, J., Viñuela, F., 1998. Principles and philosophy of modeling in biomedical research. FASEB J 12, 275–285. https://doi.org/10.1096/fasebj.12.3.275

Mirams, G.R., Pathmanathan, P., Gray, R.A., Challenor, P., Clayton, R.H., 2016. Uncertainty and variability in computational and mathematical models of cardiac physiology. J Physiol 594, 6833–6847. https://doi.org/10.1113/JP271671

Morrison, T.M., Hariharan, P., Funkhouser, C.M., Afshari, P., Goodin, M., Horner, M., 2019. Assessing Computational Model Credibility Using a Risk-Based Framework: Application to Hemolysis in Centrifugal Blood Pumps. ASAIO Journal 65, 349–360. https://doi.org/10.1097/MAT.0000000000000996

Musuamba, F.T., Bursi, R., Manolis, E., Karlsson, K., Kulesza, A., Courcelles, E., Boissel, J., Lesage, R., Crozatier, C., Voisin, E.M., Rousseau, C.F., Marchal, T., Alessandrello, R., Geris, L., 2020. Verifying and Validating Quantitative Systems Pharmacology and In Silico Models in Drug Development: Current Needs, Gaps, and Challenges. CPT Pharmacometrics Syst Pharmacol 9, 195–197. https://doi.org/10.1002/psp4.12504

Musuamba, F.T., Skottheim Rusten, I., Lesage, R., Russo, G., Bursi, R., Emili, L., Wangorsch, G., Manolis, E., Karlsson, K.E., Kulesza, A., Courcelles, E., Boissel, J.-P., Rousseau, C.F., Voisin, E.M., Alessandrello, R., Curado, N., Dall'ara, E., Rodriguez, B., Pappalardo, F., Geris, L., 2021. Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. CPT: Pharmacometrics & Systems Pharmacology 10, 804–825. https://doi.org/10.1002/psp4.12669

Nikishova, A., Veen, L., Zun, P., Hoekstra, A.G., 2019. Semi-intrusive multiscale metamodelling uncertainty quantification with application to a model of in-stent restenosis. Philos Trans A Math Phys Eng Sci 377, 20180154. https://doi.org/10.1098/rsta.2018.0154

Olberg, B., Fuchs, S., Panteli, D., Perleth, M., Busse, R., 2017. Scientific Evidence in Health Technology Assessment Reports: An In-Depth Analysis of European Assessments on High-Risk Medical Devices. Value Health 20, 1420–1426. https://doi.org/10.1016/j.jval.2017.05.011

Overgaard, R., Ingwersen, S., Tornøe, C., 2015. Establishing Good Practices for Exposure–Response Analysis of Clinical Endpoints in Drug Development. CPT Pharmacometrics Syst Pharmacol 4, 565–575. https://doi.org/10.1002/psp4.12015

Pappalardo, F., Russo, G., Tshinanu, F.M., Viceconti, M., 2019. In silico clinical trials: concepts and early adoptions. Brief. Bioinformatics 20, 1699–1708. https://doi.org/10.1093/bib/bby043

Parvinian, B., Pathmanathan, P., Daluwatte, C., Yaghouby, F., Gray, R.A., Weininger, S., Morrison, T.M., Scully, C.G., 2019. Credibility Evidence for Computational Patient Models Used in the Development of Physiological Closed-Loop Controlled Devices for Critical Care Medicine. Front Physiol 10, 220. https://doi.org/10.3389/fphys.2019.00220

Passini, E., Britton, O.J., Lu, H.R., Rohrbacher, J., Hermans, A.N., Gallacher, D.J., Greig, R.J.H., Bueno-Orovio, A., Rodriguez, B., 2017. Human In Silico Drug Trials Demonstrate Higher Accuracy than Animal Models in Predicting Clinical Pro-Arrhythmic Cardiotoxicity. Front Physiol 8, 668. https://doi.org/10.3389/fphys.2017.00668

Passini, E., Zhou, X., Trovato, C., Britton, O.J., Bueno-Orovio, A., Rodriguez, B., 2021. The virtual assay software for human in silico drug trials to augment drug cardiac testing. Journal of Computational Science, Case Studies in Translational Computer Science 52, 101202. https://doi.org/10.1016/j.jocs.2020.101202

Pathmanathan, P., Gray, R.A., 2014. Verification of computational models of cardiac electro-physiology. Int J Numer Method Biomed Eng 30, 525–544. https://doi.org/10.1002/cnm.2615

Pathmanathan, P., Gray, R.A., Romero, V.J., Morrison, T.M., 2017. Applicability Analysis of Validation Evidence for Biomedical Computational Models. Journal of Verification, Validation and Uncertainty Quantification 2. https://doi.org/10.1115/1.4037671

Pathmanathan, P., Shotwell, M.S., Gavaghan, D.J., Cordeiro, J.M., Gray, R.A., 2015. Uncertainty quantification of fast sodium current steady-state inactivation for multi-scale models of cardiac electrophysiology. Prog Biophys Mol Biol 117, 4–18. https://doi.org/10.1016/j.pbiomolbio.2015.01.008

Popper, K.R., 1994. All life is problem solving. Routledge, London.

Reuzel, R.P.B., Van Der Wilt, G.J., 2000. Health Technology Assessment and Evaluation:Back to Basics? Evaluation.

Roache, P.J., 2019. The Method of Manufactured Solutions for Code Verification, in: Beisbart, C., Saam, N.J. (Eds.), Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives, Simulation Foundations, Methods and Applications. Springer International Publishing, Cham, pp. 295–318. https://doi.org/10.1007/978-3-319-70766-2_12

Roy, C.J., Oberkampf, W.L., 2011. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. Computer Methods in Applied Mechanics and Engineering 200, 2131–2144. https://doi.org/10.1016/j.cma.2011.03.016

Rust, P., Flood, D., Mccaffery, F., 2016. Creation of an IEC 62304 compliant software development plan. Journal of Software: Evolution and Process 28. https://doi.org/10.1002/smr.1826

Sarrami-Foroushani, A., Lassila, T., Gooya, A., Geers, A.J., Frangi, A.F., 2016. Uncertainty quantification of wall shear stress in intracranial aneurysms using a data-driven statistical

model of systemic blood flow variability. J Biomech 49, 3815–3823. https://doi.org/10.1016/j.jbiomech.2016.10.005

Schaefer, J.A., Romero, V.J., Schafer, S.R., Leyde, B., Denham, C.L., 2020. Approaches for Quantifying Uncertainties in Computational Modeling for Aerospace Applications, in: AIAA Scitech 2020 Forum, AIAA SciTech Forum. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2020-1520

Simpson, S. ((Editor)), EuroScan International Network, (EuroScan), 2014. A toolkit for the identification and assessment of new and emerging health technologies (Technical Report), EuroScan Methods Toolkit. University of Birmingham, Birmingham, UK.

Smith, R.C., 2013. Uncertainty Quantification: Theory, Implementation, and Applications. SIAM - Society for Industrial and Applied Mathematics, Philadelphia.

Stahl, J.E., 2008. Modelling methods for pharmacoeconomics and health technology assessment: an overview and guide. Pharmacoeconomics 26, 131–148. https://doi.org/10.2165/00019053-200826020-00004

Strathdee, S.A., Davies, S.C., Marcelin, J.R., 2020. Confronting antimicrobial resistance beyond the COVID-19 pandemic and the 2020 US election. Lancet 396, 1050–1053. https://doi.org/10.1016/S0140-6736(20)32063-8

The Software Sustainability Institute, 2018. Checklist for a Software Management Plan. Zenodo. https://doi.org/10.5281/zenodo.1460504

Tsakalozou, E., Babiskin, A., Zhao, L., 2021. Physiologically-based pharmacokinetic modeling to support bioequivalence and approval of generic products: A case for diclofenac sodium topical gel, 1. CPT Pharmacometrics Syst Pharmacol 10, 399–411. https://doi.org/10.1002/psp4.12600

Tummers, M., Kværner, K., Sampietro-Colom, L., Siebert, M., Krahn, M., Melien, Ø., Hamerlijnck, D., Abrishami, P., Grutters, J., 2020. On the integration of early health technology assessment in the innovation process: reflections from five stakeholders. Int J Technol Assess Health Care 36, 481–485. https://doi.org/10.1017/S0266462320000756

Valerio, L.G., 2009. In silico toxicology for the pharmaceutical sciences. Toxicol Appl Pharmacol 241, 356–370. https://doi.org/10.1016/j.taap.2009.08.022

Viceconti, M, 2011. A tentative taxonomy for predictive models in relation to their falsifiability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 369, 4149–4161. https://doi.org/10.1098/rsta.2011.0227

Viceconti, M., 2011. A tentative taxonomy for predictive models in relation to their falsifiability. Philos Transact A Math Phys Eng Sci 369, 4149–61. https://doi.org/10.1098/rsta.2011.0227

Viceconti, M., Cobelli, C., Haddad, T., Himes, A., Kovatchev, B., Palmer, M., 2017. In silico assessment of biomedical products: The conundrum of rare but not so rare events in two case studies. Proc Inst Mech Eng H 231, 455–466. https://doi.org/10.1177/0954411917702931

Viceconti, M., Dall'Ara, E., 2019. From bed to bench: How in silico medicine can help ageing research. Mech Ageing Dev 177, 103–108. https://doi.org/10.1016/j.mad.2018.07.001

Viceconti, M., Emili, L., Afshari, P., Courcelles, E., Curreli, C., Famaey, N., Geris, L., Horner, M., Jori, M.C., Kulesza, A., Loewe, A., Neidlin, M., Reiterer, M., Rousseau, C.F., Russo, G.,

Sonntag, S.J., Voisin, E.M., Pappalardo, F., 2021a. Possible Contexts of Use for In Silico Trials Methodologies: A Consensus-Based Review. IEEE J Biomed Health Inform 25, 3977–3982. https://doi.org/10.1109/JBHI.2021.3090469

Viceconti, M., Henney, A., Morley-Fletcher, E., 2016. In silico clinical trials: how computer simulation will transform the biomedical industry. International Journal of Clinical Trials 3, 37–46. https://doi.org/10.18203/2349-3259.ijct20161408

Viceconti, M., Hernandez Penna, S., Dartee, W., Mazzà, C., Caulfield, B., Becker, C., Maetzler, W., Garcia-Aymerich, J., Davico, G., Rochester, L., 2020a. Toward a Regulatory Qualification of Real-World Mobility Performance Biomarkers in Parkinson's Patients Using Digital Mobility Outcomes. Sensors (Basel) 20, 5920. https://doi.org/10.3390/s20205920

Viceconti, M., Juárez, M.A., Curreli, C., Pennisi, M., Russo, G., Pappalardo, F., 2020b. Credibility of In Silico Trial Technologies—A Theoretical Framing. IEEE Journal of Biomedical and Health Informatics 24, 4–13. https://doi.org/10.1109/JBHI.2019.2949888

Viceconti, M., Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J., Musuamba Tshinanu, F., 2021b. In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. Methods, Methods on simulation in biomedicine 185, 120–127. https://doi.org/10.1016/j.ymeth.2020.01.011

Viceconti, M., Tome, M., Dartee, W., Knezevic, I., Hernandez Penna, S., Mazzà, C., Caulfield, B., Garcia-Aymerich, J., Becker, C., Maetzler, W., Troosters, T., Sharrack, B., Davico, G., Corriol-Rohou, S., Rochester, L., the Mobilise-D Consortium, 2022. On the use of wearable sensors as mobility biomarkers in the marketing authorization of new drugs: A regulatory perspective. Frontiers in Medicine 9.

Villaverde, A.F., Raimundez, E., Hasenauer, J., Banga, J.R., 2022. Assessment of Prediction Uncertainty Quantification Methods in Systems Biology. IEEE/ACM Trans Comput Biol Bioinform PP. https://doi.org/10.1109/TCBB.2022.3213914

Xin, Y., McIntosh, E., 2017. Assessment of the construct validity and responsiveness of preference-based quality of life measures in people with Parkinson's: a systematic review. Qual Life Res 26, 1–23. https://doi.org/10.1007/s11136-016-1428-x

Zarins, C.K., Taylor, C.A., Min, J.K., 2013. Computed fractional flow reserve (FFTCT) derived from coronary CT angiography. J Cardiovasc Transl Res 6, 708–714. https://doi.org/10.1007/s12265-013-9498-4

Zhao, P., Rowland, M., Huang, S.-M., 2012. Best Practice in the Use of Physiologically Based Pharmacokinetic Modeling and Simulation to Address Clinical Pharmacology Regulatory Questions. Clinical Pharmacology & Therapeutics 92, 17–20. https://doi.org/10.1038/clpt.2012.68

ANNEXES

ANNEX1: A review of the existing regulatory guidance on the use of computational models

This position paper on Good Simulation Practice does not emerge from a vacuum. In the last 20 years regulatory agencies issued guidance documents or technical standards for specific uses of In Silico methodologies. Here, we provide a short systematic review of this existing body of knowledge. The goal is not to provide details of these documents; for that it easier to consult directly to the original document; but only to provide for each a brief summary, which allows practitioners to choose which of these documents might be relevant for one's purposes.

The types of models and related documentation that are covered in this short review are related to drugs (quantitative structure-activity relationship (QSAR) models, population pharmacokinetic (Pop-PK) models, exposure-response models including the Comprehensive In Vitro Proarrhythmia Assay (CiPA) project, physiologically based pharmacokinetic (PBPK) models, and disease-drug-trial models) and to medical devices (alternative to animal testing for artificial pancreases, risk of mechanical failure, magnetic resonance imaging (MRI) safety, guidance on reporting of computational modelling studies, credibility assessment, acknowledgement in EU documentation for market access, Japanese guidelines for in silico methodologies).

It should be noted that these documents were written by different organisations, at different times, and with different purposes; thus, the language used is not consistent. However, we decided to preserve in our summaries the original texts to respect the integrity of the document.

Drugs - Quantitative structure-activity relationship (QSAR) models

Quantitative structure-activity relationship models (QSAR models) are classification or regression models. In drug discovery they are used to identify molecular structures with low non-specific activity and good inhibitory effects of specific targets; they are also used to estimate the octanol-water partition coefficient (logP), important information to evaluate how a substance behaves with respect to factors like bioavailability (druglikeness).

The only piece of regulatory guidance available for this type of model is the following:

ENV/JM/MONO(2004)24. Report from the Expert Group on (Quantitative) Structure-Activity Relationships on the Principles for the Validation of (Q)SARs. Paris, France: Organization for Economic Co-operation and Development (OECD) Expert Group on QSARs.

https://read.oecd.org/10.1787/9789264085442-en?format=pdf

Quantitative structure–activity relationship (QSAR) models are regression or classification models used in the chemical and biological sciences and engineering. There are two types of QSAR models, regression or classification QSAR models. Like other regression models, QSAR regression models relate a set of "predictor" variables (X) to the response (potency) variable (Y), while classification QSAR models relate the predictor variables to a categorical value of the response variable. In QSAR modelling, the predictors consist of physicochemical properties or theoretical molecular descriptors of chemicals; the QSAR response-variable could be a biological activity/potency of the chemicals. QSAR models are first developed based on a dataset of chemicals to describe the relationship between chemical structures and biological activity. Then, QSAR models can be used to predict the activities

of new chemicals. QSAR models can be as simple as a statistical regression, involve molecular dynamics calculations (e.g., 3D-QSAR based on binding affinity), or more complex and advanced models such as machine learning models. QSAR models rarely include a mechanistic model of the physiology beyond the molecular scale: they capture either the mechanistic chemistry of the drug action at the molecular scale or build phenomenological relations with clinical endpoints.

The OECD Principles for Validation of (Q)SAR for regulatory consideration are:

- a defined endpoint
- an unambiguous algorithm
- a defined domain of applicability
- appropriate measures of goodness-of–fit, robustness and predictivity
- a mechanistic interpretation, if possible

Drugs - Pop-PK

Pharmacokinetics (PK) investigates how a drug is absorbed, distributed, metabolized, and eliminated from the body. Population pharmacokinetics models (popPK) are informed by concentration-time data from multiple individuals frequently pooled across multiple studies, and are used to for allometric scaling, exposure-response, bioequivalence, and many uses.

The pieces of regulatory guidance available for this type of models are:

- FDA-2019-D-2398 (CDER, CBER). Population Pharmacokinetics Guidance for Industry
- EMA/CHMP/EWP/185990/06. Committee for Medicinal Products for Human Use (CHMP). Guideline on reporting the results of population pharmacokinetic analysis.

FDA-2019-D-2398 (CDER, CBER). Population Pharmacokinetics Guidance for Industry

https://www.fda.gov/media/128793/download

This FDA guidance focus on how to conduct a Pop-PK analysis but contains limited information regarding models' validation. The section on model validation states: "Model validation depends on the objective of the analysis and should follow a fit-for-purpose approach". Most of the recommendation refer to the type of plots that could be used to present the validation results, such as goodness-of-fit (GOF) plots, dependent variable versus the individual predictions plots, etc.

EMA/CHMP/EWP/185990/06. Committee for Medicinal Products for Human Use (CHMP). Guideline on reporting the results of population pharmacokinetic analysis.

https://www.ema.europa.eu/en/reporting-results-population-pharmacokinetic-analyses#current-effective-version-section

Population pharmacokinetics (Pop-PK) is the study of variability in drug concentrations between individuals (healthy volunteers or patients). It comprises the assessment of variability within the population, associated with patient characteristics such as age, renal function, or disease state. The non-linear mixed effects modelling approach has become increasingly used for Pop-PK. The EMA

"Guideline on reporting the results of population pharmacokinetic analyses" assume such approach is used. In contrast to the FDA guidance on Pop-PK analyses, this guideline does not provide guidance on how to conduct a Pop-PK analysis, but rather provides guidance on points to consider when develop the analysis plan and the final analysis report.

The analysis plan should at least include:

- the objective(s) of the analysis
- a brief description of the study (or studies) from which the data originates
- the nature of the data to be analysed (how many subjects, rich or sparsely sampled)
- the procedures for handling missing data and outlying data
- the general modelling aspects (e.g., software, estimation methods, diagnostics)
- the overall modelling procedure/strategy
- the structural models to be tested (if this has been decided)
- the variability models to be tested
- the covariates and covariate models to be tested together with a rationale for testing these covariates based on, for example, biological, pharmacological and/or clinical plausibility.
- the algorithms/methods to be used for covariate model building
- the criteria to be used for the selection of models during model building and the inclusion of covariates (e.g., objective function value, level of statistical significance, the goodness of fit plots, standard error, inter-individual variability, clinical relevance)
- The model evaluation/qualification procedures to be used.

The final report should include the following sections:

- Summary
- Introduction
- Objectives
- Data
- Methods
- Results
- Discussion

Drugs – dose-response models

Whereas pharmacokinetics models predict how of the drug reaches the target, pharmacodynamics models predict the effect that the drug will produce on the target biological system. An important category of pharmacodynamics models are dose-response models (also known as exposure-response models).

The concept of exposure and response are not always unequivocally defined. The broad term exposure is used to refer to dose (amount of a drug enters into the body) as well as to various measures

of acute or integrated drug concentrations in plasma and other biological fluids. Similarly, response refers to a direct measure of the pharmacologic effect of the drug. Response measures include a broad range of endpoints or biomarkers.

EMA has only guidelines for specific products (e.g., EMA/CHMP/594085/2015, which targets microbials), whereas FDA has a general guidance.

FDA (CDER, CBER) 2003. Guidance for Industry: Exposure-Response Relationships — Study Design, Data Analysis, and Regulatory Applications.

https://www.fda.gov/media/71277/download

Points to Consider in Study Design of Exposure-Response Analysis:

- Crossover, fixed dose, dose response
- For immediate, acute, reversible responses
- Provide both population mean and individual exposure-response information
- Safety information obscured by time effects, tolerance, etc.
- Treatment by period interactions and carryover effects are possible; dropouts are difficult to deal with
- Changes in baseline-comparability between periods can be a problem
- Parallel, fixed dose, dose response
- For long-term, chronic responses, or responses that are not quickly reversible
- Provides only population mean, no individual dose response
- Should have a relatively large number of subjects (one dose per patient)
- Gives good information on safety
- Titration
- Provide population mean and individual exposure-response curves, if appropriately analysed
- Confounds time and dose effects, a particular problem for safety assessment
- Concentration-controlled, fixed dose, parallel, or crossover
- Directly provides group concentration-response curves (and individual curves, if crossover) and handles inter-subject variability in pharmacokinetics at the study design level rather than data analysis level
- Requires real-time assay availability

In the process of PK-PD modelling, it is important to describe the following prospectively:

- Statement of the Problem
- Statement of Assumptions
- Selection of the Model
- Validation of the Model

- There are also recommendations on the structure of the reporting:
- The response variable and all covariate information
- An explanation of how they were obtained
- A description of the sampling design used to collect the PK and PD measures
- A description of the covariates, including their distributions and, where appropriate, the accuracy and precision with which the responses were measured
- Data quality control and editing procedures
- A detailed description of the criteria and procedures for model building and reduction, including exploratory data analysis.

Drugs - Extrapolation models

EMA/189724/2018. Reflection paper on the use of extrapolation in the development of medicines for paediatrics

https://www.ema.europa.eu/en/extrapolation-efficacy-safety-paediatric-medicine-development#current-version-section

Extrapolation is defined as 'extending information and conclusions available from studies in one or more subgroups of the patient population (source population(s)), or in related conditions or with related medicinal products, in order to make inferences for another subgroup of the population (target population), or condition or product, thus reducing the amount of, or general need for, additional evidence generation (types of studies, design modifications, number of patients required) needed to reach conclusions.' While the focus of discussion here is on extrapolation for the development of medicines in children, the underlying principles may be extended to other areas.

Extrapolation Concept

Existing information about the disease, the drug pharmacology and the clinical response to treatment should be collated across studies and target populations. Factors that might impact the effects of treatment from different studies and target populations should be identified.

The primary focus will usually be to establish a line of reasoning about the relationship between exposure and clinical responses. Where data are available to establish that a relationship (e.g., exposure-response) in the target population is similar to the study population the knowledge gained from the study population can be incorporated into the extrapolation concept and will not need to be addressed in the extrapolation plan.

For other relationships or factors, reliable and informative data might not be available. These gaps in knowledge give rise to assumptions in the extrapolation concept that need to be investigated in the extrapolation plan before the extrapolated effects of treatment in the target population can be considered as a sound basis for regulatory decision-making.

Where possible, quantitative methods should be used for the collation of available data and the investigation of potential modifiers of the treatment effect. A structured extrapolation plan should be provided.

Extrapolation Plan

The gaps in knowledge and the assumptions identified in the extrapolation concept determine the objectives(s) and methodological approaches for the tests and trials that need to be conducted to draw inferences that are relevant for the target population. These tests and trials should be conducted to generate evidence that strengthens and ultimately, based on success criteria, confirms the extrapolation concept. Specifically, the extrapolation plan will address whether regulatory decisions can rely on the initial, or revised, expectations on the effects of treatment in the target population, or if more data need to be generated.

Extrapolation plans will differ according to the extent of assumptions in the extrapolation concept. Data in the study population might establish that there are so few important modifiers of the treatment effect that clinical outcome can be predicted through similarity in drug exposure or in the magnitude of PD response. Alternatively, data from the source population might be limited such that the influence of one or more factors needs to be investigated through generation of some additional clinical data from the target population. The extreme case would be where gaps in knowledge might be such that extrapolation is not a viable approach.

Mitigation of uncertainty

Whilst conclusions from an extrapolation approach can give a sound basis for regulatory decision making, the data generated may not be sufficient to address all uncertainties related to a specific research question for the target population. For example, an acceptable degree of patient benefit on short-term efficacy outcomes, sufficient to support authorisation, might be established based on an extrapolation approach, but quantification of how this effect translates into longer-term outcomes might not be available. When there is a well-reasoned scientific uncertainty to be addressed to enhance the understanding of the effect of treatment with implications for better labelling and better use in clinical practice, the extrapolation plan can continue post-authorisation to reduce the identified uncertainty.

Drugs - PBPK

As mentioned above, whereas pharmacokinetics models predict how of the drug reaches the target, pharmacodynamics models predict the effect that the drug will produce on the target biological system. While PK models were historically developed without any mechanistic assumptions, by simply fitting experimental data with statistical models, Physiologically Based Pharmaco-Kinetics (PBPK) models predict the absorption, distribution, metabolism and excretion of a drug by relying on the mechanistic knowledge that anatomy, physiology, physics and chemistry can provide. In most cased PBPK models are so-called *grey-box models*, in the sense they are built combining mechanistic and empirical (e.g., data-driven) knowledge.

EMA/CHMP/458101/2016. Guideline on the reporting of physiologically based pharmacokinetic (PBPK) modelling and simulation.

https://www.ema.europa.eu/en/reporting-physiologically-based-pharmacokinetic-pbpk-modelling-simulation#current-effective-version-section

The guideline recommends the essential information that needs to be reported when reporting PBPK modelling and simulation studies:

- Objective and regulatory purpose
- Background information
- Qualification
- Model parameters
- Assumptions
- System-dependent parameters
- Drug parameters and the drug model
- Model development
- Simulation of the intended scenario
- Platform and drug model evaluation
- Sensitivity analyses
- Evaluation of the predictive performance of the drug model
- Results
- Discussion of the regulatory application

In particular, with respect to model evaluation, the EMA wrote: "A comprehensive summary of the system and drug model evaluation should be provided. A thorough evaluation of the drug model is important if the model is to be used to simulate novel situations, e.g., drug interaction or PK in a different population. An evaluation of the model should be presented in appendix with sufficient detail in the report to support confidence for regulators in the application of the model in their decision-making.

The appendix should provide some additional recommendations:

- The validation should include the investigational drug PBPK model (treatment model).
- Validation studies should be done against human experiments with multiple doses.
- Simulation should be performed on populations of interest of at least 100 subjects.

The guidance suggests the types of plots to be used to compare predictions to experiments.

"The acceptance criteria (adequacy of prediction) for the closeness of the comparison of simulated and observed data depends on the regulatory impact and needs to be considered separately for each application."

FDA/CDER/2018. Physiologically Based Pharmacokinetic Analyses — Format and Content Guidance for Industry.

https://www.fda.gov/media/101469/download

The FDA guidance covers:

- Overview of Modeling Strategy
- Modeling Parameters
- Simulation Design

- Electronic Files and Other Documentation
- Software
- Model Verification and Modification
- Model Application

The introduction section should provide:

1. a high-level synopsis of the drug's physicochemical, PK, and PD properties;
2. the exposure-response relationships for the efficacy and safety of the drug to the extent that they are known;
3. a brief PBPK-related regulatory history (i.e., prior interactions with the FDA and other regulatory agencies) to provide context for the PBPK analyses;
4. cross-referencing to PBPK study reports previously submitted to the FDA for different intended uses at different stages of the development of the same drug substance or the same drug product.

The Materials and Methods section should include "**sufficient information to allow FDA reviewers to duplicate and evaluate the submitted modelling and simulation results and to conduct supplemental analyses when necessary.**"

Electronic files related to modelling software and simulations should be submitted along with the PBPK study report.

FDA/CDER draft guidance October 2020. The Use of Physiologically Based Pharmacokinetic Analyses —Biopharmaceutics Applications for Oral Drug Product Development, Manufacturing Changes, and Controls - Guidance for Industry

https://www.fda.gov/media/142500/download

This guideline covers the concept of quality by design (QbD) principles and propose that the application of PBPK modelling could be expanded to pharmaceutical drug product development, manufacturing changes, and controls. It is applicable to oral formulations, only.

In addition to the general considerations (which follow a similar structure as in the previously described guideline), specific applications of PBPK modelling to support product quality are described:

1. Development of Clinically Relevant Dissolution Specifications (Method and Acceptance Criteria):
    a. Aid in Biopredictive Dissolution Method development
    b. Support Clinically Relevant Dissolution Acceptance Criteria
2. Establishment of Clinically Relevant Drug Product Quality Specifications (Other Than Dissolution)
3. Quality Risk Assessment for Pre- and Post-approval Changes and Risk-Based Biowaivers

Drug - Chemicals - PBK

OECD Guidance document on the characterization validation and reporting Physiologically Based Kinetic (PBK) models for regulatory purposes. Adopted April 2021[83].

http://www.oecd.org/chemicalsafety/risk-assessment/guidance-document-on-the-characterisation-validation-and-reporting-of-physiologically-based-kinetic-models-for-regulatory-purposes.pdf

This OECD guidance supports the use of PBK models for chemical risk assessment as an alternative to animal testing. PBK models are the same of PBPK models; the change in terminology is due to the fact that OECD targets not only drugs development, but also the safety assessment of chemical products, so the terminology is more generic.

It describes the key steps for characterizing and validating such a model to improve model credibility and communication between modelers and regulators, but it does not provide good practices for model development. Interestingly, this guidance accounts for the fact that PBK models are most often calibrated with in vitro or in silico data, as in vivo kinetic data may not be available. However, this guidance explicitly states that goodness-of-fit and predictivity for PBK models requires *in vivo* kinetic data without stating their prospective or retrospective character. It also discusses the validation as a term that may be understood differently by model developers and regulators.

First, a regular PBK modelling workflow is described. Notably, it includes a step called model performance, which covers model validation, sensitivity, variability and uncertainty analyses, and predictive capacity.

The regulatory assessment part explains what is considered in the validation of a PBK model, taking into account the CoU. It provides two tools: (1) a model reporting template for model developers and (2) an evaluation checklist of model applicability for regulators.

The recommendations from the template for reporting are the following:

- Name of model
- Model developer and contact details
- Summary of model characterization, development, validation, and regulatory applicability
- Model characterization (following the steps of the aforementioned modelling workflow)
- Identification of uncertainties
- Model implementation details
- Peer engagement (input/review)
- Parameter tables
- References and background information

The checklist for regulators is split into a context/implementation section and an assessment of validity section. The later covers the biological basis of the model, the theoretical basis of model equations, reliability of input parameters, uncertainty & sensitivity analysis and goodness-of-fit & predictivity.

83 https://www.oecd.org/chemicalsafety/risk-assessment/guidance-document-on-the-characterisation-validation-and-reporting-of-physiologically-based-kinetic-models-for-regulatory-purposes.pdf

CiPA: Comprehensive In Vitro Proarrhythmia Assay

While the CiPA project has not produced guidelines yet, it is worth mentioning it here. The project, a collaboration between various regulatory agencies including FDA and EMA, aims to define a new approach to evaluate the risk that a new drug may cause *torsades de pointes* (TdP), an abnormal heart rhythm that can lead to sudden cardiac death, based on a suite of in vitro assays coupled to in silico models of cardiac electrophysiologic activity. The project will provide data to ICH to update the S7B and E14 guidances. The most recent is this below, where some guidelines for in silico models is provided.

ICH E14/S7B Implementation Working Group: Clinical and Nonclinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential Questions and Answers. Draft version, Endorsed 27 August 2020

https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-e14/s7b-clinical-nonclinical-evaluation-qt/qtc-interval-prolongation-proarrhythmic-potential-questions-answers-step-2b_en.pdf

"The following general principles should be applied to all proarrhythmia risk prediction models intended to be used as part of an integrated risk assessment for regulatory purposes. While the main focus of these principles is to evaluate a model's predictivity of TdP risk, they are general enough to guide the development of models predicting different types of proarrhythmia.

1. A defined endpoint consistent with the context of use of the model.
2. A defined scope and limitations of the model. This includes the experimental protocols to generate model input (experimental data capturing pharmacological effect of drug), and the compounds tested should have the same arrhythmic mechanisms covered by the model.
3. A prespecified analysis plan and criteria to assess model predictivity. The analysis plan should include methods to separate the training and validation steps. In the training step, a series of reference compounds is used to adjust the model. In the validation step, another series of reference compounds is used to evaluate the performance of the pre-specified model. The reference compounds used for the training and validation steps should not overlap.
4. A fully disclosed algorithm to translate experimental measurements (model input) to proarrhythmia risk (model output), allowing independent reproduction of the model development process using the associated training and validation datasets to re-evaluate the model performance.
5. The uncertainty in the model inputs should be captured and propagated to the model predictions. The experimental variability associated with model input should be quantified using appropriate statistical methods and then translated into probabilities of the predicted risk.
6. A mechanistic interpretation of the model, which describes the relationship between the model inputs and mechanism for the arrhythmia."

CM&S for Medical devices – alternative to animal testing for artificial pancreas

Kovatchev BP, Breton MD, Dalla Man C, Cobelli C. In silico model and computer simulation environment approximating the human glucose/insulin utilization. Food and Drug Administration Master File MAF 1521. 2008

https://moodle.adaptland.it/pluginfile.php/20224/mod_data/content/42094/In-Silico%20Application%20-%20paper%201.pdf

In January 2008, a computer simulator of type 1 diabetes mellitus was accepted by the FDA Center for Devices and Radiological Health (CDRH) as a substitute for animal trials for the preclinical testing of control strategies in artificial pancreas studies[84]. Soon after, a first investigational device exemption was granted by the FDA for a closed-loop control clinical trial on the basis of results from this simulation tool. To the authors' knowledge, this is the first instance of a regulatory decision on medical devices where a computer model prediction is accepted as replacement of an in vivo experiment.

Medical devices – risk of mechanical failure

ASTM F2514-08, Standard Guide for Finite Element Analysis (FEA) of Metallic Vascular Stents Subjected to Uniform Radial Loading, ASTM International, West Conshohocken, PA, 2008, www.astm.org

The American Society for Testing and Materials (ASTM) F2514-8 was the first technical standard related to the use of a physics-based model to predict the performance of medical devices. According to ASTM, the purpose of the guide is to "establish recommendations and considerations for the development, verification, validation, and reporting of structural finite element models used in the evaluation of the performance of a metallic vascular stent design undergoing uniform radial loading. This standard guide does not directly apply to non-metallic or absorbable stents, though many aspects of it may be applicable. The purpose of a structural analysis of a stent is to determine quantities such as the displacements, stresses, and strains within a device resulting from external loading, such as crimping or during the catheter loading process, and in-vivo processes, such as expansion and pulsatile loading".

Published in 2008, the standard establishes general requirements and considerations for using Finite Element Analysis techniques for the numerical simulation of metallic stents subjected to uniform radial loading.

The basic idea was that for highly standardised experimental bench tests (for stents the ASTM F2477 – 07: Standard Test Methods for in vitro Pulsatile Durability Testing of Vascular Stents), a finite element model could reliably predict the outcome of the experiment, and thus be used to reduce and refine, and even in some low-risk cases, replace the bench experiment itself.

While the experimental tests being supplemented were extremely simple, the adoption of this early standard introduced in the regulatory space the concept that a model prediction could be used in place of an experiment within the regulatory process.

[84] https://doi.org/10.1177/193229680900300106

FDA-2020-D-0957 Non-Clinical Engineering Tests and Recommended Labeling for Intravascular Stents and Associated Delivery Systems – Guidance for Industry and FDA Staff

https://www.fda.gov/media/71639/download

This is not a guideline specifically focused on computational methods, but it includes some non-binding recommendations for the use of modelling in the pre-clinical assessment of stents:

- You should establish protocols for all experiments or computational analyses, including acceptance criteria when applicable, before you perform the tests
- We recommend that you determine the stress- strain response, endurance limit, and post-processing mechanical properties through physical experiments or computational models that simulate stent material properties, manufacturing, and deployment processes.
- FDA recommends that you include the following elements in your stress/strain analysis and test report for each stent design.
    - Computational Model and Inputs
        - We recommend that you clearly identify and explain the sources and values of all inputs and assumptions used to create the stress/strain analysis model. You should identify any software used for analysis. We recommend that finite element analysis reports include the element types used to model the stent, loading surfaces, and boundary conditions. We also recommend that you indicate if mesh refinement analysis was performed and clearly describe how you model the surrounding vessel/tissue and the type of contact elements used. Specifically, we recommend that you consider the following:
    - Model Geometry
        - We recommend that you clearly describe the stent and vessel geometry used. If symmetry is used, we recommend that you explain why this is appropriate for your model.
        - If you do not model all of your stent sizes, we recommend that you explain why the modelled stent size is the worst case with respect to critical stresses. We recommend that you address the effect of dimensional variation within allowable tolerances on the results of the stress/strain analysis (i.e., maximum critical stress).
        - We recommend that you provide a justification for the physiological relevance of your vessel model parameters (e.g., vessel compliance).
    - Type of Element & Mesh Refinement Analysis
        - We recommend that you specify the number and type of elements used in your mesh, including any mesh refinement in transition regions or regions of complex geometry.
        - We recommend that you perform a mesh refinement analysis to ensure that the solution is independent of element size. If you do not believe mesh refinement analysis is necessary for your model, we recommend that you provide a justification for not conducting such an analysis.

- Contact Elements
    - We recommend that you specify the type of contact defined between any 2 contacting bodies modelled in your analysis; e.g., the vessel and outer surface of your stent.
    - Material Properties (Constitutive Model)
    - We recommend that you clearly describe the material stress/strain behaviour of your stent in graphical and equation form. This discussion should include, but is not limited to the following considerations:
        - Linear vs. non-linear
        - Isotropic vs. anisotropic
        - Temperature-dependent behaviour of raw vs. processed material.
- Finite Element Analysis (FEA) Validation
    - We recommend that you validate your FEA (material properties, geometry, and boundary conditions) with experimental bench testing. For example, you could perform radial loading of your device and compare the force-displacement results with FEA of a simulated radial loading experiment.

FDA-2019-D-1261. Technical Considerations for Non-Clinical Assessment of Medical Devices Containing Nitinol

https://www.fda.gov/media/123272/download

This guidance also includes some recommendations for the use of computational models:

- 2. Computational Stress/Strain Analyses
    - If you plan to conduct computational analyses, we recommend the following to ensure the unique thermomechanical properties of nitinol are properly captured:
        - a. The constitutive laws applicable to nitinol can differ substantially from traditional metals. Therefore, you should simulate nitinol material with an appropriate material model. You should document and justify the parameters used in the material model.
        - b. Material model parameters can be obtained from ASTM F2516 “Standard Test Method for Tension Testing of Nickel-Titanium Superelastic Materials.” Test specimens should be representative of the final manufactured device (e.g., including heat treatment and surface processing steps). Testing should be conducted at a temperature representative of the clinical use environment (e.g., 37°C for implantable devices).
        - c. Your computational analysis should include the effect of any shape setting steps in your manufacturing process since these will relieve pre-existing stresses.
        - d. If your device is subjected to cyclic loading during use, we recommend that you calculate fatigue safety factor(s) using a constant life curve. Unlike

traditional metals, which utilize stress-based fatigue life estimates (e.g., Goodman, Soderberg diagrams), using a constant life mean versus alternating strain diagram has been found to provide a good model for fatigue life prediction for nitinol.34 Fatigue life of nitinol is sensitive to composition and processing. Therefore, we recommend that you generate a constant life curve specific to your device by experimental testing of nitinol samples that are representative of your final manufactured device (e.g., including heat treatment and surface processing) rather than leveraging data not specific to your device. Since fatigue life can be adversely or favourably affected by pre-strain (e.g., from crimping of a stent onto a delivery catheter), we recommend you consider and discuss the effects of pre-strain. We recommend that you state and justify the method used to calculate mean and alternating strain for fatigue safety factors (e.g., scalar or tensor).

- e. You should validate the computational model used to analyse the nitinol device, and justify the validation activity relative to the context of use (CoU) of the computational model, the risk and role of the computational model in decision-making, and the range of conditions assessed relative to those in the CoU. We also recommend that you justify your choice of the parameter measured (e.g., force, strain) and loading path in your validation activities.

- We recommend that the submission of computational stress/strain analysis reports follow the "Reporting of Computational Modeling Studies in Medical Device Submissions Guidance".

ASTM F2996-13, Standard Practice for Finite Element Analysis (FEA) of Non-Modular Metallic Orthopaedic Hip Femoral Stems, ASTM International, West Conshohocken, PA, 2013, www.astm.org

The other widely used bench test for medical devices was the fatigue testing of hip stems. The late 1970s saw a number of fatigue fractures for various hip stems, which drove the ISO to the development of a technical standard for the execution of bench tests in the late 80s (ISO 7206-3:1988). Every regulatory agency quickly required these bench tests to provide the marketing authorisation for new hip stem designs. The 2013 ASTM F2996 standard provides a computational complement to the ISO7206 fatigue tests.

ASTM F3161-16, Standard Test Method for Finite Element Analysis (FEA) of Metallic Orthopaedic Total Knee Femoral Components under Closing Conditions, ASTM International, West Conshohocken, PA, 2016, www.astm.org

ASTM F3334-19, Standard Practice for Finite Element Analysis (FEA) of Metallic Orthopaedic Total Knee Tibial Components, ASTM International, West Conshohocken, PA, 2019, www.astm.org

Extension to two popular bench tests for orthopaedic implants: the fatigue testing of the femoral component and of the tibial component of knee replacements.

ASTM WK64097. New Practice for Spinal Fusion Cage Computational Modeling

This ASTM work item (standard under development) should provide the computational counterpart to the "ASTM F2077-18: Test Methods For Intervertebral Body Fusion Devices" experimental protocol.

Medical devices - Guidance on reporting modelling studies

FDA-2013-D-1530. Reporting of Computational Modeling Studies in Medical Device Submissions. Guidance for Industry and Food and Drug Administration Staff.

In 2013 the FDA CDRH started to work on a guidance document on how to report the results of computational modelling studies in medical device regulatory submissions. The first draft was published in 2014, with the final version issued in 2016. Already in 2017, 220 of the 1500 new medical devices submitted to the FDA included computer models and simulations as evidence in their regulatory submissions (Source: AABME).

According to this guidance such report should include:

- Executive Report Summary
- Background/Introduction
- Code Verification
- System Configuration
- System Properties
- System Conditions
- System Discretization
- Numerical Implementation
- Validation
- Results
- Limitations

The document also provides in annex more detailed instructions for specific type of models:

- Computational Fluid Dynamics and Mass Transport
- Computational Solid Mechanics
- Computational Electromagnetics and Optics
- Computational Ultrasound
- Computational Heat Transfer

FDA-2021-D-0980. Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions.

https://www.fda.gov/media/154985/download

This is to date (July 2022) the most complete guidance available on this topic; the linked version is the final draft of the revision published in December 2021, which will replace the 2016 version. The guidance provides a generalized framework for assessing credibility of computational modelling in a regulatory submission. Since the document is entirely relevant trying to make a summary here would be difficult. Instead, we provide an extract of the index, for the core chapters:

A. Preliminary steps

Question of Interest

Context of use (CoU)

Model risk

B. Credibility Evidence

Code verification results

Model calibration evidence

General non-CoU evidence

Evidence generated using bench-top conditions to support the current CoU

Evidence generated using in vivo conditions to support the current CoU

Evidence generated using bench-top conditions to support a different CoU

Evidence generated using in vivo conditions to support a different CoU

Population-based evidence

Emergent model behavior

Model plausibility

C. Credibility Factors and Credibility Goals

D. Adequacy Assessment

Medical devices - Credibility assessment

ASME VV40:2018. Assessing Credibility of Computational Modeling through Verification and Validation: Application to Medical Devices

In 2018 was published the official version of the first technical standard that specifies a systematic approach to the credibility assessment for computational models, which results inform a regulatory submission on medical devices. ASME reports that the standard "determine and justify the appropriate level of credibility for using a computational model to inform a decision"; thus, its scope is not limited to the regulatory purpose. Since this standard is one of the centrepieces of this position paper, we provide here only a summary.

The standard introduces a general principle: the credibility of a computational model should be commensurate with the risk associated in using the model to influence a decision. The concept of risk-based acceptability is very general and has been rapidly endorsed also for other medical products, as evidenced by a 2019 paper providing a concrete example for a medical device[85], a 2020 paper proposing its use also in drug development for PBPK models[86], and a concept reiterated in a 2021 paper where the proposed use is extended to any mechanistic model used in drug development[87].

[85] https://doi.org/10.1097/mat.0000000000000996
[86] https://doi.org/10.1002/psp4.12479
[87] https://doi.org/10.1002/psp4.12669

FDA-2003. Knee Joint Patellofemorotibial and Femorotibial Metal/Polymer Porous-Coated Uncemented Prostheses - Class II Special Controls Guidance Document for Industry and FDA

https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-radiation-emitting-products/knee-joint-patellofemorotibial-and-femorotibial-metalpolymer-porous-coated-uncemented-prostheses

the guidance mentions the use of finite element analysis to support regulatory submissions: "...Alternatively, finite element analysis (FEA) or other calculations with validation of the model and assumed values may be appropriate".

Medical devices: EU side

EU-MDR-2017/745: Medical Device Regulation.

This new legislation defines the rules concerning the placing on the market, making available on the market, or putting into service of medical devices for human use and accessories for such devices in the EU. While the legislation does not include specific requirements for the use of computer models and simulation for the development of medical devices, it explicitly acknowledges their use:

- "where appropriate, the results of biophysical or modelling research the validity of which has been demonstrated beforehand" (Annex I, 10.1a))
- "the pre-clinical testing, for example laboratory testing, simulated use testing, computer modelling, the use of animal models" (Annex VII, 4.5.4.a)).

Japanese guidelines

Ministry of Economics, Industries, and Japan Agency for Medical Research and Development. Guidelines for Developing in silico evaluation. March 2019.

We were able to access an English translation of this document, dated 2019, that targets specifically In Silico Clinical Trials for medical devices. The guideline suggests for the model evaluation:

1. Construct a scenario based on a mathematical model (i.e., an expectation based on an experiment that such a phenomenon will be observed if the model is correct).
2. Run all or part of the scenario as an experiment.
3. Consider whether the experimental results support the scenario.
4. Consider whether the numerical model is reasonable in light of established theories and techniques.
   (a) Consider whether the numerical model (experimental system) including numerical methods adequately treats the system subject to the mathematical model.
5. We will discuss the hypotheses that mathematical models entail and the conditions under which they are valid.
   (a) Consider whether their logic is interconnected and whether there are any holes.
   (b) When multiple hypotheses are included, it may not be possible to isolate them depending on the conditions of the experiment.

The document also contemplates a case they call V&V of Unknown Provenance (VOUP), in which methods and others are known and commonly available, but adequate recordings of "who did the experiment" and "how" are unavailable. This resembles the concept Software of Unknown Provenance in International Electrotechnical Commission (IEC) 62304. They make a parallel with experimental methods, where some "good practice" must be adopted for the experiment to be valid, but it is not always possible to trace back when such good practice was adopted and validated. Similarly, in silico models might use practices that are known to be valid, but that cannot be documented back to the original developer. When this is the case, the guideline recommends that "the degree of VOUP should be clarified, and the in silico evaluation should proceed based on the items that can be accepted although the details are unknown."

The guideline discusses some essential steps:

1. Determination of the subject. This is what we call Context of Use.
2. Setting targets for achieving the task. Objectives set based on available knowledge and standards.
3. List of components to be calculated.
4. Description of the physical phenomena of numerical calculations and mathematical models representing them
5. Various parameter settings in the mathematical model. These include:
    a. The shape and dimension of the object to be numerically calculated.
    b. Boundary conditions for numerical calculations
    c. Initial conditions for numerical calculation
    d. External input to the object required to perform numerical calculations (e.g., energy, load, force, etc.).
    e. Characteristic values that appear in mathematical models and are essential for numerical computation.
    f. Unit system of numerical calculation used for shape, dimension, characteristic value, external input, etc., and check the consistency of units.
6. Description of the numerical method
7. Description of numerical results
8. Confirmation of conformity with numerical calculations
9. Validation of numerical calculations
10. Validation of in silico assessment

10.7.1. Chinese guidelines

We were able to access an English translation of this document, dated 20, that targets model informed drug development. The guideline suggests the following points to be addressed during the "Implementation of model analysis"

1. Quality control

2. Model assumption
3. Model verification
4. Model-based analysis plan
1. Model-based analysis report

of which 1-3 could be understood of dealing with the evaluation of models.