Building the European Virtual Human Twin

Call: Accelerating best use of technologies (DIGITAL-2021-DEPLOY-01)
Work program year: DIGITAL-2021-2022
Topic: ID DIGITAL-2021-DEPLOY-01-TWINS-HEALTH
Grant Agreement No: 101083771

Public consultation version of the first draft of the VHT Roadmap

http://www.doi.org/10.5281/zenodo.8070381

Document open for public consultation mid June – July 15th

EDITH is a Coordination and Support Action funded by the Digital Europe program of the European Commission under the grant agreement n. 101083771.

Foreword

The first draft of the VHT roadmap needs to be delivered by the 31st of July 2023. The final roadmap is due at the end of the EDITH CSA (September 2024). The present document is a preliminary version of this first draft. It shows the envisioned structure and contains the main points that will be included in the first draft of the roadmap. The level of detail and refinement in the description of these points is still variable and is continuously being refined and elaborated. However, we are making this version already publicly available for comments and feedback so this input can still be taken incorporated in the first draft of the roadmap submitted to the European Commission.

This first draft aims to **capture the main lines** of thinking and identify the research, infrastructure and other challenges that need to be discussed in the remainder of the EDITH project but will require further development and support beyond EDITH. For the technology, standards, regulatory and legal elements the draft gives an overview of the state of the art and an analysis of the VHT-specific needs. **No definitive choices have been proposed** in this document.

Throughout the document, the **blue highlighted** text is text that serves to guide the reader of this document explaining what will be put in specific section when it is not yet ready or is requesting specific inputs. The document has been made available as a **google doc with commenting rights** for everyone. This is done so everyone is able to keep track of suggestions that are being made (and by whom – in case further clarifications are needed). For general discussions, we encourage everyone to use the slack channel on the **In Silico World Community of Practice (ISW_CoP)**. This will facilitate a much more lively discussion than comment balloons on a text file.

As a final note: we are very happy to receive all your comments, additions, suggestions and criticisms – but we are equally interested in learning which elements you are particularly happy or in agreement with.

Link to ISW_CoP: https://insilico.world/community/join-the-community-of-practice-channels/

Table of Contents

List of Figures

List of Tables

Acronyms (to be updated)

Acronym	Full name
AI	Artificial Intelligence
ASME	American Society of Mechanical Engineers
ATMP	Advanced Therapy Medicinal Product
B2B	Business-to-Business
BBCT	Bologna Biomechanical CT
BMD	Bone Mineral Density
CBER	Center for Biologics Evaluation and Research
CDRH	Center for Devices and Radiological Health
CE	Conformité Européenne
CoP	Community of Practice
CoU	Context of Use
CRO	Contract Research Organisation
CSA	Coordination and Support Action
DLT	Distributed Ledger Technology
DOP	Data Object Pose
DOT	Data Object Type
DT	Digital Twin
DTH	Digital Twin in Healthcare
DXA	Dual-energy X-ray absorptiometry
EC	European Commission
EMA	European Medicine Agency
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reproducible
FDA	Food and Drug Administration
FFR	Fractional Flow Reserve
GA	Grant Agreement
GDPR	General Data Protection Regulation
GPU	Graphic Processing Unit
HPC	High performance computing
HTA	Health Technology Assessment
ICT	Information and Communications Technology
ID	Identity
IHI	Innovative Health Initiative
IPR	Intellectual Property Rights
IST	In Silico Trials
ISW	In Silico World
ML	Machine Learning
MRI	Magnetic Resonance Imaging
nD	n-dimensional
NGO	Non-governmental Organisation
OEM	Original Equipment Manufacturer
PC	Project Coordinator
PM	Project Manager
QoI	Quantity of Interest
SaaS	Software as a Service
SaMD	Software as a Medical Device
SME	Small and Medium-sized Enterprises
SOP	Standard Operating Procedure
VHT	Virtual Human Twin
VV-40	Verification and Validation 40

WP	Work Package
yo	Years old

1 Introduction

The Virtual Human Twin (VHT) is an integrated multi-scale, multi-time, and multi-discipline representation of quantitative human physiology and pathology. Its realisation through collaborative distributed knowledge and resource platforms is specifically designed to accelerate the development, integration, and adoption of patient-specific predictive computer models, which will be used as clinical decision support systems for personal health forecasting or as methodologies for the development and de-risking of personalised medical products. The **vision of EDITH** is to facilitate the realisation of the opportunities presented by VHTs for the benefit of patients, healthcare providers, regulatory bodies and industry, and other ecosystem actors, both within Europe and globally.
EDITH is a Coordination and Support Action (CSA) funded by the European Commission, capitalising on the developments of digital technologies, employment of high-performance computing, availability and access to research and healthcare data in Europe, with the **mission** of creating a roadmap to go from the currently available resources (which often focus on a single organ or a single function system) to a data-driven and knowledge-driven fully integrated multi-scale and multi-organ whole-body VHT. EDITH will facilitate this process by building an evolutionary ecosystem driven by a consensus among the relevant European communities and implemented through the aid of advanced technological tools, such as a data/model repository and a simulation platform.

The objectives of the EDITH project are the following.

- To frame an ecosystem of digital twins in healthcare within the EU. EDITH is conducting a **mapping** of actors, initiatives, resources, and barriers in the digital twins, with the aim of ensuring adequate clinical representation and fostering the integration of all relevant stakeholders such as developers, technology and infrastructure providers, end-users, regulatory agencies, and Health Technology Assessment (HTA) bodies.
- To build a **roadmap** towards an integrated Virtual Human Twin (VHT), identify the main research challenges and infrastructure needs and formulate clear policy recommendations. It will also address interoperability, computability and health information integration, identifying implementation needs/barriers and developing a strategy for the clinical deployment of the VHT model and its uptake in personalised clinical decision-making.
- To develop a **federated and cloud-based repository** of digital twins in healthcare (data, models, algorithms, and good practices), pooling together existing resources across Europe and providing access to relevant existing data and model repositories. The ecosystem will be leveraged to create a repository catalogue with available resources and recruit resources from the consortium and beyond.
- To outline a **simulation platform** supporting the transition towards an integrated VHT implemented as a public infrastructure, providing a one-stop shop to design, develop, test, and validate single-organ digital twins and combine them with others for the integrated VHT models. Five use cases (cancer, cardiovascular, intensive care, osteoporosis, and brain) have been pre-selected to be developed as prototypes to show the added value of a simulation platform.

This document is a preliminary version of the first draft of the VHT roadmap (due 31/7/2023). It consolidates the results of a wide range of discussions, consultations and expert inputs. This version is made publicly available now and will remain open for until 15/7/2023 in order to capture additional comments, questions and inputs defines from the entire ecosystem. The final VHT roadmap is due in September 2024.

*This **first draft aims to capture the main lines of thinking and identify the research, infrastructure and other challenges that need to be discussed in the remainder of the EDITH project but will require further development and support beyond EDITH.** For the technology, standards, regulatory and legal elements the draft gives an overview of the state of the art and an analysis of the VHT-specific needs. **No definitive choices have been proposed in this document**.*

This vision is based on the current state of art and information in a highly dynamical field. It is hence expected that new aspects and technologies will lead to extension of the given views.

2 Genesis of the vision and roadmap outline

This document results from work that has been initiated in the grant preparation phase and has continued through the review phase and the project execution phase until official submission at DATE. Given the pressure of the first version of the deliverable to be ready by Month 10 (end of July 2023), the work has been carried out mainly within the scientific community of experts with proven track records in the addressed scientific disciplines and fields, with input from specific external experts including industrial and clinical colleagues in the EDITH Industry Advisory Board and other expert meetings.

2.1 Meetings to discuss vision and roadmap

The main discussion meetings are recurring online meetings with the consortium, the industry advisory board, public discussion meetings as well as meetings with various communities of the ecosystem. This was complemented with a number of on-site consortium meetings and a deep thinkers meeting in Rome.

- October 11th 2022: EDITH kick-off meeting
- Since October 2022 (ongoing): 4 Working groups (Mapping, Vision, Repository/Platform, Sustainability), each meeting on a biweekly, later partially weekly, basis with dedicated agenda for each working group.
- November 29-30 2022: EDITH consortium meeting Leuven (1)
- Since December 2023 (ongoing): Industry Advisory Board meeting every 2 weeks
- January 30-31 2023: EDITH consortium meeting Leuven (2)
- May 16-17 2023: Deep Thinkers Meeting Rome (100 experts covering all stakeholders groups)

Additionally, several public events and community of practice meetings (*e.g.* Avicenna Alliance, VPHi) have taken place where EDITH consortium members have been invited to present the current status of the developing VHT vision. All those inputs have been included into the current document.

2.2 Writing of the roadmap

The written representation of the aforementioned meetings and discussions resulted in the current document. A first public document detailing the vision and roadmap outline was published and opened for comments and discussion[1]. Those inputs have been included in the current document. Finally, a manuscript summarising the vision articulated below is available in the ArXiv[2].
The organisation of the following phase with public discussions and community writing, is discussed at the end of the document (Section 9).

[1] EDITH CSA Deliverable 3.1: Vision for the Virtual Human Twin and Roadmap Outline (2023). https://doi.org/10.5281/zenodo.7796845
[2] Viceconti *et al*. (2023); arXiv:2304.06678v1. https://doi.org/10.48550/arXiv.2304.06678

3 State of the Art and Maturity of Virtual Human Twin elements

3.1 Introduction

This chapter aims to provide a state of the art is some of the building blocks and technologies of the VHT. These have been kept very brief but additional information can be easily found in the scientific literature and online. The chapter starts with an brief discussion of the state of the art in In Silico Medicine, highlighting both the data-driven (AI) and knowledge-driven (mechanistic) perspectives. This is followed by an overview of wearables and the development of data-driven twins with a real-time connection to sensor inputs.

Note on the nomenclature Virtual Human Twin – Digital Twin in Healthcare: *The term Virtual Human Twin is used in the context of this document to indicate the concept of combining resources towards the development of a fully integrated Digital Twin for personalised health and care. In addition, it is also used as the brand name of the overarching initiative as well as the infrastructure to be developed. In the context of this initiative, Digital Twins represent smaller entities focusing on a specific applications. Digital twins are virtual representations of a physical object, process or system across its life-cycle. It uses data and other sources (that in the context of healthcare applications do not necessarily need to be real-time) to enable learning, reasoning, and dynamically recalibrating for monitoring, diagnostics and prognostics.*

3.2 In silico medicine

In general, a predictive model of human (patho)physiology is built on observations (data) with variable amounts of prior mechanistic facts or hypotheses, in the following summarised as "knowledge". We call data-driven models those containing no mechanistic knowledge, and mechanistic models those which, in an ideal case, are fully based on mechanistic knowledge. However, these are the extremes of the *in silico* spectrum, with most actual models situated somewhere in between[3]. Lately, the boundaries have been further blurring with explainable Artificial Intelligence (AI), physics informed Scientific Machine Learning (ML) or ML-based surrogates of mechanistic models, all emerging as techniques that combine the best of both worlds and usually in some way result in a superior solution. What modelling approach is preferred in each situation depends on the question of interest, the Context of Use (CoU) of the model, and the available data and knowledge. The model resources in the VHT will be covering the entire **spectrum from data-driven to knowledge-driven**. The specific resource used in the simulations will depend on the question of interest, the available data, the computational resources and other factors and dependencies.

In silico medicine has achieved substantial progress in the last decade. Single organ subsystem DTs are **entering the marke**t and **regulatory science** is in full development with several ASME and ISO **standards** in place.

3.2.1 AI in health

We use the term **AI for a set of methods** that, **based on data and for a given set of objectives, generate outputs** such as content, predictions, recommendations, or decisions influencing the environments they interact with. These methods include amongst others ML approaches, logic- and knowledge-based approaches, or statistical approaches. With the overall purpose being to provide essential explanatory accounts of the healthy and diseased human body and to facilitate actionable model-based predictions on the effect of interventions on the health and disease of individuals, there is a key role for AI in the vision of the VHT. AI DTs are entering the market at record speed and are the subject of many policy and strategy initiatives as discussed in chapter 6. One reason is that it is able to identify relations in large bodies of data that had so far not been recognized.

We foresee that **AI technologies can be employed in at least five ways in building the VHT**. The first is to help increase the sheer number of systems that can be assessed, from *e.g.* heart models based on image processing to the assessment of numbers of candidate compounds that may be used for drug

[3] WHITE PAPER: the role of Artificial Intelligence within in silico medicine (2022). https://doi.org/10.5281/zenodo.8064147

treatments. The second way is in the development and incorporation of surrogate models or ML based PDE solvers which can replace some features of multiscale models in order that more expensive and computation-intensive aspects of these simulations are replaced by a more efficient approach; such models will also play a key role in uncertainty quantification studies, which are essential for implementing and adopting actionable models that can be used in *e.g.* clinical decision support. The third way is in helping to parametrise and personalise mechanistic models, *e.g.* using advanced Bayesian inverse uncertainty quantification techniques and relying on increasing and enriching multimodal individual health and clinical data. The fourth way is completely data driven predictive models that would *e.g.* provide decision support on treatment options, or patient specific prognosis prediction. Such models need to be explainable, *e.g.* by integrating prior mechanistic knowledge. A fifth way could be the generation of hypotheses from data, that may inform and guide the development of mechanistic multilevel models.

In addition to its involvement in the creation of DTs and the VHT, AI will also be used for natural language processing (NLP), relying amongst others on large language models such as chatGPT. Using NLP, we can *e.g.* automatically **extract information from the literature** (for mapping and building knowledge graphs and models). Knowledge graphs can be produced that **provide feedback to the VHT platform users** on resources that can be coupled, as well as specific data needs that would allow to generate a more comprehensive DHT *etc.* Finally, NLP can produce narratives of the platform outcomes by automatically generating reports that contain the required technical or clinical information, tailored for the targeted user, *e.g.* a clinician.

3.2.2 Mechanistic models in health

Mechanistic models have been the more familiar approach from classical engineering, biological, physical or chemical sciences. Not surprising, most of the models addressing health-related questions in the last decades were **based on conceptual analogies** to problems solved within these classical quantitative disciplines. The evaluation of their achievement, despite being significant if compared to the situation before this endeavour has been undertaken, suffers considerably from the originally unrealistic high expectations. Formulating body function in terms of mechanisms means using the spelling out of **body physiology and pathophysiology** at each level and scale in its natural alphabet of molecules, cells, sub-organ units, whole organs and the whole body, eventually leading to a multiscale, multilevel DT of human in health and disease. This has the advantage, that, if the mechanistically DTH is accurate, it is able to correctly predict the outcome of the so far unknown interventions, for which no or only very little data is available. On the other hand, to arrive at an accurate DTH accurate data on all parameters along the mechanistic chains represented within the DTH simultaneously are required as well as iterations between the forming DTH and validation experiments to evaluate intermediate DTH stages. Over the last decades, partially driven by mechanistic models and the associated raised questions, experimental modalities have been considerably advanced so that accurate data acquisition is now largely possible. **Mechanistically-based DTH development** and **experimental modality development** as well as **experimental design** go hand in hand, guiding and fertilising each other. Experience accumulated over decades in responding to biological and clinical questions has led to a large body of DTs and has advanced our understanding on how such DTs should be built. Within a network of collaborations between experimentalists, clinicians, and DTs developers, and flanked by AI with the potential to significantly accelerate mechanistic hypothesis building, mechanistic-based DTH components, entire DTH, and finally the VHT are in reach.

3.3 Wearables & data-driven twins

Wearables form real-time data generators Edge AI technologies capable of supporting remote monitoring of patients and mass screening populations towards disease prediction and progression monitoring. In this context, the value of **wearable technologies** relies on real-time data collection, non-invasiveness and capacity to support multi-model embodiments in very small form factors. While the majority of technological efforts to date mostly concern physiological monitoring of vital signs[4] and of activity[5] (*e.g.* cardiac monitoring, core body temperature, blood pressure *etc.*), there is a stringent need

[4] Soon S *et al.*, BMJ Innov. 2020; 6:55–71. http://dx.doi.org/10.1136/bmjinnov-2019-000354
[5] Saoutieff E *et al.*, Sensors 2021; 21(5):1802. https://doi.org/10.3390/s21051802

of advanced monitoring technologies for specific biomarkers to take into account the most relevant dynamics of disease symptoms and nature, such as, for instance, the detection of cytokine storm in COVID 19 or sepsis patients[6]. The VHT platform will be designed to collect and exploit such real-time data as input to future generations of DTHs under high security and privacy data constraints.
Interest in **point-of-care** and large-scale bio-chemical sensing has triggered a lot of research in the past decade into **highly scalable technologies** like ISFETs[7] and CHEMFETs[8] that could benefit from the scaling and low power consumption of semiconductors chips. Nevertheless, numerous challenges remain related to such devices aiming at ion, metabolite, protein and hormone sensing in biofluids as well as to their readout electronics, in terms of sensitivity into sub-nM range, selectivity and integrated multi-parameter sensing. Recent years have seen an increased research effort in semiconductor sensors benefitting from new 1D and 2D semiconducting nanostructures[9], as well as new readout architectures and topologies. **Next-generation Edge AI** systems for healthcare are evolving towards hybrid systems such as coarse-grained reconfigurable arrays (CGRAs), for a flexible and energy-efficient acceleration of a wide range of AI-based embedded bio-signals processing kernels. **Federated learning** enables the training of ML models for DT[10] over distributed training data while respecting privacy, without the training data ever leaving its device of origin.
Finally, another important aspect for advancing the state-of-the-art in wearable devices is the **biofluid of interest** for non-invasive or minimally invasive approaches as compared to the blood, the medical gold standard . While sweat[11] appears in many recent reports as fluid of interest for non-invasive sensing, the correlations of markers in sweat with blood is so far poorly established and contamination is still an issue. Despite a semi-invasive access, today the interstitial fluid seems to hold higher promise for biosensing to capture real-time biomarker dynamic for building DT due to less dilution of biomarkers and high correlations of marker concentrations and dynamics with blood (serum)[12]. In addition, there is also ongoing research in breath and saliva analysis with the target of replacing the invasive blood examination with breath/saliva biosensors, but to date these are still in an early phase.

3.4 Infrastructures & platforms

In the context of European computational infrastructures for scientific research, over the recent years there has been a trend towards federation of resources provided by multiple academic institutions across national borders. Today, repositories and platforms come in many shapes and sizes; there is a large number of Research Infrastructures and the number is growing fast due to the role of science as a driver of innovation and economic growth, the ever-increasing importance of information technology in science, and the benefits that accrue from greater collaboration and scale. There are a lot of **research infrastructures** (RIs) that support medical sciences. The vast majority[13] are "single-site" (small range of services, limited users), while about only one third of the RIs listed are "distributed", with facilities spread across multiple locations. On the other hand, in the 2021 roadmap[14] of European Strategy Forum on Research Infrastructures (ESFRI), there are 16 RIs in the Health and Food category and 4 in Data, Computing and Digital Research Infrastructures, all 20 of them of distributed type. This occurs against the background of ongoing migration towards exascale-level systems, coupled with new management and architectural approaches, focusing on modular, integrated and lightweight solutions, which enable end-users across all scientific domains to use these future **supercomputing systems[15]**. There is also ongoing work on addressing organisational and legislative issues involved in sharing and processing scientific data in federated environments, as evidenced *e.g.* by the European Health Research and Innovation Cloud[16]. Below, a non-exhaustive overview is provided, highlighting some initiatives being

[6] Yang L *et al.*, Sig Transduct Target Ther 2021; 6:255. https://doi.org/10.1038/s41392-021-00679-0
[7] Lee C-S *et al.*, Sensors 2009; 9(9):7111–7131. https://doi.org/10.3390/s90907111
[8] Capua L *et al.*, IEEE Trans. Electron Devices 2022. https://doi.org/10.1109/TED.2022.3144108 .
[9] Bolotsky A *et al.*, ACS Nano 2019; 13(9):9781–9810. https://doi.org/10.1021/acsnano.9b03632
[10] Rieke N *et al.*, NPJ Digit. Med. 2020; 3:119. https://doi.org/10.1038/s41746-020-00323-1 .
[11] Brasier N, Eckstein J, Digit Biomark 2019; 3:155–165. https://doi.org/10.1159/000504387
[12] Kim Y, Prausnitz MR, Nat Biomed Eng 2021;5:3–5. https://doi.org/10.1038/s41551-020-00679-5
[13] https://portal.meril.eu/meril/
[14] https://www.esfri.eu/esfri-roadmap-2021
[15] Kranzlmüller D, Höb M, Computing and Informatics 2021;39(4):617–621. https://doi.org/10.31577/cai_2020_4_617
[16] Aarestrup FM *et al.*, Genome Med 2020;12:18. https://doi.org/10.1186/s13073-020-0713-z

highly relevant to the VHT. Synergistic and strategic collaborations with these initiatives will allow to identify more effectively the unmet needs and services complementing those offered elsewhere.
A key example of the trend towards exascale is the **EuroHPC** initiative[17], which supports development of several petascale and pre-exascale systems across Europe, such as for instance **LUMI**, a major pre-exascale system under the EuroHPC umbrella. Another key initiative on advanced computing and data management (more bottom-up science-driven, complementing the Euro-HPC top down approach) is **PRACE**[18]. Additionally, **Elixir**[19] is a sustainable distributed European infrastructure that brings together life science resources and where scientists can access biological data, software, cloud, storage and supercomputers from a single infrastructure. In addition, many national and regional resource federations, such as the **PL-Grid** infrastructure[20], augment and extend the European-level efforts, providing additional opportunities related to academic studies and research data processing.
In this HPC context there are several centers of excellence focused on biomedical applications: **CompBioMed**[21], the Centre of Excellence focused on the use and development of computational methods for biomedical applications, and **PerMedCoE**[22], the HPC/Exascale Centre of Excellence for Personalised Medicine in Europe. For instance, PerMedCoE's tools observatory and community benchmark will provide interesting input when discussing the multiscale integration of resources and the model validation in the creation of the HDT.
FENIX research infrastructure[23] has emerged as a major provider of **federated supercomputing and storage resources**. The distinguishing characteristic of the Fenix e-infrastructure is that data repositories and scalable supercomputing systems are in close proximity and well integrated. FENIX has been developed in the Human Brain Project (HBP) to enable computation and data intensive research. It is linked to **EBRAINS RI, a digital research infrastructure selected to the ESFRI roadmap**[24]. The HBP is developing the concept of the human brain twin as an enabler for future brain research and medicine,
EOSC (European Open Science Cloud)[25] will provide a federated environment for European scientists to access, find and share FAIR data and services for storage, computation and analysis for researchers in various scientific fields. The **Gaia-X**[26] is a European initiative that will create a **federated and secure data infrastructure, allowing c**ompanies and citizens to collate and share data, keeping control over them and retaining data sovereignty, within a networked system linking many cloud service providers together. **OpenAIRE**[27] is a partnership of more than 50 institutions, aiming to implement and establish effective Open Access and Open Science policies. OpenAIRE's distributed Research Information Infrastructure is responsible for the management, analysis, provision, cross-linking and sharing of all research outcomes. We can also mention **HealthyCloud**[28] that aims at providing a distributed interoperable ecosystem for health research.
The **EUDAT Collaborative Data Infrastructure (EUDAT CDI)**[29] is one of the largest infrastructures of integrated data services and resources supporting research in Europe. EUDAT envisions sharing and preserving data across borders and disciplines by enabling data stewardship within and between European research communities through a Collaborative Data Infrastructure (CDI), a common model and service infrastructure for managing data spanning all European research data centres and community data repositories. EUDAT offers heterogeneous research data management services and storage resources, supporting multiple research communities as well as individuals, through a geographically distributed, resilient network and data is stored alongside supercomputers.

[17] EuroHPC portal, https://eurohpc-ju.europa.eu/
[18] https://prace-ri.eu/
[19] https://elixir-europe.org/
[20] PL-Grid infrastructure, https://www.plgrid.pl/en
[21] https://www.compbiomed.eu/
[22] https://www.permedcoe.eu/
[23] FENIX research infrastructure https://fenix-ri.eu/about-fenix
[24] https://ebrains.eu/
[25] https://eosc-portal.eu/
[26] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html
[27] https://www.openaire.eu/
[28] https://healthycloud.eu/
[29] https://www.eudat.eu/

The recently started **EUropean Federation for CAncer Images project (EUCAIM)**[30] aims to deploy a pan-European digital federated infrastructure of FAIR cancer-related de-identified images from Real-World. EUCAIM infrastructure will preserve the data sovereignty of providers and provide a platform for developing and benchmarking AI tools. EUCAIM focuses on the field of cancer, includes an Atlas of Cancer Images, pathology, molecular and laboratory data and will address the fragmentation and cluster the existing cancer image repositories by exploiting the AI4HI initiative[31], European Research infrastructures, regional and national repositories.

3.5 Scientific organisations / Clinical institutions / Industry

Results of the mapping will be summarized here. Focus will be on organisations that represent the different communities in the ecosystem (VPH, COMBINE, AA, DIGITALEUROPE, MedTechEurope, COCIR, EFPIA, HOPE, EUHA,…).

3.6 Regulatory, HTA and standardisation actors

There is a growing interest in in silico methodologies among **Health Technology Assessment (HTA)** agencies, regulators and standardisation bodies. In this section we provide a brief overview of the key initiatives and the connections between these entities and the use of predictive computer models in healthcare.

3.6.1 HTA actors

According to the internationally accepted new definition[32], HTA is a multidisciplinary process that uses explicit methods to determine the value of a health technology at various stages of its lifecycle. In this context, the term "health technology" refers to a broad range of generic interventions, including tests, devices, medicines, vaccines, procedures, programs, or systems developed to prevent, diagnose, or treat medical conditions. An in silico methodology can be thus considered at the same time a “health technology” and an “health technology assessment tool”[33].

The first scenario is related to the **use of computer modelling and simulations as a clinical decision support system**. One example is given by the HeartFlow service[34] that uses patient-specific models generated from computed tomographic coronary angiography (CTCA) images to predict the Fractional Flow Reserve (FFR), a measurement of the pressure difference across a specific segment of a coronary artery that provides information in support of the functional evaluation of the appropriate treatment for patients with suspected coronary artery disease. HeartFlow was the first application to receive FDA marketing authorization as a software-based medical device.

The second scenario that considers the **in silico methodology as an “health technology assessment tool”** is associated with the use of computer modelling and simulation as a source of evidence when developing a new medical product or evaluating its safety, efficacy and cost- effectiveness. Numerous examples in the literature demonstrate the potential use of computer modelling and simulation in the discovery, design and pre-clinical stage to reduce the costs of innovation and the time to market for biomedical products. A list of possible examples on the use of in silico trials to reduce, refine and replace clinical experimentations can be found in[35] []. One notable success is the UVA/Padova Type 1 Diabetes Simulator[36] which was initially accepted by the FDA in 2008 as an alternative to pre-clinical animal experimentation to test insulin treatments, including artificial pancreas systems. In silico methodologies can contribute during the clinical development phase by providing additional evidence through computer simulations, thereby reducing the cost and time of the clinical trials and the number of patients involved. A valuable application in this context is the use of In Silico Augmented Clinical Trials, where virtual patients are considered to explore and understand the effects of medical interventions on less common phenotypes, which may be challenging to enroll in traditional clinical

[30] https://eucanimage.eu/
[31] Kondylakis H *et al*. Eur Radiol Exp (2022); 6: 29. https://www.doi.org/10.1186/s41747-022-00281-1
[32] O’Rourke B *et al*., Int. J. Technol. Assess. Health Care 2020; 36(3):187–190, https://doi.org/10.1017/S0266462320000215
[33] https://insilico.world/sito/wp-content/uploads/2023/05/Position-Paper-GSP-R6.pdf
[34] Nørgaard BL *et al*., J Am Coll Cardiol 2014; 63:1145-55. https://doi.org/10.1016/j.jacc.2013.11.043
[35] Viceconti M *et al*. IEEE J Biomed Health Inform. (2021); 25(10):3977-3982. https://www.doi.org/10.1109/JBHI.2021.3090469.
[36] Visentin R *et al*..J Diabetes Sci Technol. (2018); 12(2):273-281. http://www.doi.org/10.1177/1932296818757747 .

studies[37]. In silico methodologies can offer added value also during the post-approval of medical product, aiding in optimizing real-world data collection and understanding cost-effectiveness mostly through data-driven approaches[38].
Several HTA agencies have demonstrated interest in silico methods and have proposed related initiatives. Here are some representative examples:

- National Institute for Health and Care Excellence (NICE) - United Kingdom: NICE has expressed interest in in silico methods and their potential impact on HTA. They have highlighted the use of modelling and simulation techniques to support decision-making, particularly in areas such as medical devices and diagnostics. They presented a 5-year strategic plan where the use of digital health technologies is among the 6 key identified trends[39].
- French National Authority for Health (HAS) - France: HAS has been exploring ways to integrate modelling and simulation techniques into their assessment processes and has developed a classification system for digital solutions used in healthcare based on three criteria: intended use, capacity to provide a personalized response, and autonomy in decision-making[40].
- Acquas (Agency for Health Quality and Assessment of Catalonia) is an organization based in Catalonia, Spain, dedicated to assessing the quality and effectiveness of healthcare interventions and technologies. Acquas conducts HTA studies to inform decision-making processes related to healthcare policy, resource allocation, and reimbursement decisions. The agency plays a significant role in evaluating the clinical and economic aspects of healthcare technologies, including pharmaceuticals and medical devices. Acquas recently contributed to the discussion on the role of artificial intelligence within in silico medicine[41].

At the international level, one of the most important organizations dedicated to health technology assessment is the **Health Technology Assessment International (HTAi)[42]**. HTAi brings together experts, professionals, and organizations involved in HTA from around the world. It serves as a global platform for knowledge exchange, collaboration, and development of HTA methodologies and practices. HTAi organizes annual conferences and provides opportunities for networking and collaboration among HTA stakeholders. They have organized workshops and sessions to discuss the role of in silico trials to complement or supplement randomised controlled trials of new technologies.
At the EU level, **EUnetHTA[43]** was established to create an effective, collaborative and sustainable network for HTA across Europe – to help develop reliable, timely, transparent, and transferable information to contribute to HTA in European countries.

3.6.2 Regulatory actors

Regulatory agencies play a crucial role in providing guidance, establishing standards, and promoting the safe and effective use of computational modelling and simulation in healthcare.

The **US Food and Drug Administration (FDA)** has been actively involved in exploring the use of in silico methodologies and computational modelling in regulatory decision-making. They have developed guidance documents and initiatives such as:

- FDA 2016 guideline on Reporting of Computational Modelling Studies in Medical Device Submissions[44]
- FDA 2018 guidance on PBPK models[45]
- FDA draft guidance document [46] that outlines a generalised framework for assessing model credibility that relies heavily upon the ASME VV-40 standard[47].

The **European Medicines Agency (EMA)** has also recognized the potential of in silico methodologies and computational modelling. They have established the "Innovation Task Force" to facilitate the

[37] Viceconti M *et al*. Int J Clin Trials (2016); 3(2):37-46. : http://dx.doi.org/10.18203/2349-3259.ijct20161408
[38] Courcelles *et al*. Front. Med. Technol. (2022); 4:8103157. https://doi.org/10.3389/fmedt.2022.810315
[39] https://static.nice.org.uk/NICE%20strategy%202021%20to%202026%20-%20Dynamic,%20Collaborative,%20Excellent.pdf
[40] https://www.has-sante.fr/upload/docs/application/pdf/2019-07/rapport_analyse_prospective_20191.pdf
[41] Geris L et al. VPH Institute & Avicenna Alliance White Paper (2022). https://www.doi.org/10.5281/zenodo.8064147
[42] https://htai.org/
[43] https://www.eunethta.eu/
[44] https://www.fda.gov/media/87586/download
[45] https://www.fda.gov/media/101469/download
[46] https://www.fda.gov/media/154985/download
[47] https://www.asme.org/codes-standards/find-codes-standards/v-v-40-assessing-credibility-computational-modelling-verification-validation-application-medical-devices

uptake of innovative methods in the development of medicines. They published some guidance documents on the topic such as:

- EMA guideline on reporting PBPK models[48]
- Guidance for "Qualification of novel methodologies for medicine development"[49]

Different regulatory pathways and strategies for the acceptance of in silico methodologies in medical device and drug development can be identified, including:

- **Certification of a SaMD**: Nowadays, regulatory authorities widely acknowledge software designed for medical purposes as a distinct category of medical devices known as Software as a Medical Device (SaMD). Both the FDA's Center for Devices and Radiological Health (CDRH) and the European Union's CE-marking process have established regulatory pathways specifically designed for these types of technologies. Notably, there is a particular focus on SaMDs with predictive capabilities. An example class of SaMD includes the HeartFlow service solution, discussed in the previous section.
- **Qualification of in silico methodologies**: The FDA and EMA offer qualification pathways for medical device and drug development tools. While the FDA provides qualification for both, the EMA currently only provides it for drug development tools. The process for qualifying a new methodology is not mandatory but highly recommended; it involves requesting qualification advice from the regulatory authority followed by a formal request for qualification opinion. If a positive qualification opinion is obtained and there are no criticisms from the experts, the developer can utilize the methodology to generate evidence in a marketing authorization application for a new medical product.

3.6.3 Standardisation actors

The standardisation actors are either grass-root standardisation initiatives or official committees of **standard defining organisations** (SDOs), which define technical or clinical standards. The most important SDOs for technical standards are the **International Standards Organisation** (ISO) with its Technical Committees 215 (Health Informatics) and 276 (Biotechnology) and the **CEN / CENELEC** with its technical committee 251 (Health Informatics). In addition, there are several other institutions, societies and authorities defining standards relevant for VHTs. Beside that there are also some **grass-root communities** like COMBINE, the DIN, the Avicenna Alliance and the Genome Alliance for Genomics and Health (GA4GH). A full list of SDOs defining technical standards is given in Table 1 in the annex.

There are also several institutes, organisations, and initiatives which define clinical standards. One of the most prominent of them is the Clinical Data Interchange Standards Consortium (CDISC). A comprehensive list of SDOs relevant for defining clinical standards is given in Table 2 in the annex.

3.7 Public policy at EU level

3.7.1 Legal framework

There is a growing body of EU law that is relevant for Virtual Human Twins. We have already had in place the regulation governing medical device software (**Medical Device Regulation**, Regulation (EU) 2017/745, enacted 2017) and data protection (**GDPR** (EU) 2016/679, enacted 2016, in force since 2018). More recently, following the launch of the European strategy for data (COM/2020/66 final) in early 2020, the European Commission has proposed a number of regulations aiming to create an enabling environment for health data use and research, and to build the infrastructure to do so. These regulations include:

- **Data Governance Act** proposed in November 2022, and entering into force in September 2023 (Regulation (EU) 2022/868).
- **AI Act** proposed in April 2021 (COM/2021/206 final), currently in trilogue negotiations
- **Data Act** proposed in February 2022 (COM/2022/68 final)
- **European Health Data Space** regulation proposed in May 2022 (COM(2022) 197 final)

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

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

A more detailed discussion of the impact of these regulations on the Virtual Human Twin can be found in section 6.4.

3.7.2 EU funding instruments

The European Commission has been actively funding many successful initiatives in the area of simulation and modelling in healthcare towards the realisation of digital twins in healthcare (*cfr* section 3.8). To foster a thriving ecosystem around digital twins in healthcare, various public policy measures and funding instruments need to be implemented such as research grants to support the development and testing of digital twin technologies, infrastructure investment, collaborative initiatives between public and private stakeholders, training and education programmes, pilot projects and testbeds, data sharing and interoperability, and incentives for adoption. Current funding for the development of digital twins in healthcare can be accessed under three main funding instruments in the field of healthcare, research and innovation established by the EC in 2021:

- **The Digital Europe Programme** established via Regulation (EU) 2021/694[50]
- **Horizon Europe** — the framework programme for research and innovation, laying down its rules for participation and dissemination, established via Regulation (EU) 2021/695[51]
- **EU4 Health Programme** - established via Regulation (EU) 2021/522[52]

The **Digital Europe Programme** is designed to bring digital technology to businesses, citizens, and public administrations, while aiming to improve the EU's competitiveness in the global digital economy, contribute to bridging the digital divide across the EU and strengthen Europe's competences in digital technology through large-scale distribution. Digital Europe Programme is running from January 2021 to December 2027, with an expected budget of €7.5 billion, under direct management of the European commission, and performs its work towards five specific goals: Achieving high-performance computing, Artificial Intelligence, Honing advanced digital skills, Optimising digital interoperability and Advancing Cybersecurity. A network of European Digital Innovation Hubs provides technological expertise for the participating entities, private and public. This initiative is co-financed by the Member States and is open to non-EU countries that are part of the EEA, candidate countries, European Neighbourhood Policy countries, and other countries under agreement.
The recently released Digital Europe Work Programme for the period 2023-2024 describes under Cloud, Data and Artificial Intelligence, the development of the Platform for advanced virtual human twin (VHT) models (2.3.4). The objective of this action is to develop a distributed platform that provides access to a federated repository of virtual human twins (VHTs) related resources, open source software toolkits, and computational services for developing, testing, and integrating VHT models. The platform will be used for personalised care, medical training, surgical intervention planning, and professional training and educational purposes. It will provide controlled and secure access to an environment of simulation and visualisation tools, open access and proprietary data, and model assets for advanced modelling. The platform will be interoperable with augmented and virtual reality environments and will be based on access to computational services enabled by strategic digital capabilities with links to suitable testing and experimentation facilities. The indicative global budget is €20 million and the indicative duration of the action is 24-36 months[53].

Horizon Europe[54] is the key funding programme for Research and Innovation, acting for the period 2021 to 2027, with a total budget of €95.5 billion. Its overarching goal is to propel scientific, technological, economic, and societal advancement in the EU by investing in Research and Innovation. This programme intends to maximise Union added value by focusing on milestones that can only be achieved by the Member States acting in cooperation. The programme has several objectives, which include promoting scientific intelligence, supporting the creation and application of high-quality knowledge, skills, technologies, and solutions, providing training for researchers, and supporting new

[50] https://eur-lex.europa.eu/EN/legal-content/summary/digital-europe-programme-2021-2027.html
[51] https://eur-lex.europa.eu/eli/reg/2021/695/oj
[52] https://eur-lex.europa.eu/EN/legal-content/summary/eu4health-programme-2021-2027.html
[53] https://ec.europa.eu/newsroom/dae/redirection/document/94609
[54] https://research-and-innovation.ec.europa.eu/system/files/2022-06/ec_rtd_he-investing-to-shape-our-future_0.pdf

talent. The program also aims to strengthen the impact of research and innovation in implementing Union policies, promote innovation in European industry and societal challenges, foster all facets of innovation, facilitate technology transfer, and encourage excellence-based participation from all member states to strengthen the attractiveness of the European Research Era. The first Strategic Plan covers the years 2021 – 2024.
The specific programme implementing Horizon Europe includes three pillars:

- Pillar I-Excellent Science, reinforcing and extending the excellence of the Union's science base (European Research Council expected budget €16 billion, Marie Sklodowska-Curie Actions - €66 billion and Research infrastructure -2.4 €billion).
- Pillar II – Global challenges & European Industrial Competitiveness, boosting key technologies and solutions underpinning the EU policies & Sustainable Development Goals. Expected funding for the health cluster is €8.246 billion.
- Pillar III – Innovative Europe, stimulating market-creating breakthroughs and ecosystems conducive to innovation, with expected budgets of 10.6 billion, including up to €527 million for European innovation ecosystems, and €3 billion for the European Institute of Innovation and Technology (EIT).

There are five Research and Innovation missions areas in Horizon Europe, which includes the mission on Cancer, targeting by 2030: more than 3 million more lives saved, living longer and better, achieve a thorough understanding of cancer, prevent what is preventable, optimise diagnosis and treatment, support the quality of life of all people exposed to cancer, and ensure equitable access to the above across Europe.

EU4Health Programme
The EU4Health Programme is the EU's response to the COVID-19 pandemic and has an expected budget of €5.3 billion during the 2021-27 period. It aims to support and complement national policies to improve and protect human health, strengthen health systems, and increase resource efficiency within the EU. In this light, this programme has 10 objectives:

- Disease prevention and boosting of health campaigns.
- Promotion of international health initiatives
- Efficient prevention and response to cross-border health threats
- Complementing national stockpiling of essential crisis-relevant products
- Creating a reserve of healthcare and support staff
- Improving availability and accessibility of medical products and devices
- Strengthening digital tools and health data storage and transformation
- Increasing access to healthcare
- Developing and implementing EU health legislation
- Collaboration between Member States' healthcare systems

3.8 VHT initiatives & actors in Europe and at member state level

Results of the mapping will be summarized here

[We invite all interested parties to share information about regional or national initiatives (centers, grass roots initiatives, funding programs, policy priorities, …) that promote the development of digital twins in healthcare. A short paragraph or weblink is appreciated.]

Name of the initative	Country	Contract person (from within the community)	description	link

3.9 Conclusions

TBD

4 Vision for the virtual human twin

4.1 Introduction

TBD

4.2 Vision and barriers

4.2.1 Introduction

The EC has recently launched the *Destination Earth* initiative. "Destination Earth (DestinE) aims to develop – on a global scale - a highly accurate digital model of the Earth to monitor and predict the interaction between natural phenomena and human activities. As part of the European Commission's Green Deal and Digital Strategy, DestinE will contribute to achieving the objectives of the twin transition, green and digital". The vision of developing a comprehensive digital twin of the planet comes from the need to define policies that can guide an incredibly complex system, the ecosphere, into a state more desirable for the human species.

There is a strong parallel with human health. The need to define policies that guide the health of the citizens of the European Union towards improving healthcare delivery and the quality of life for each person clashes with the difficulty of making predictions for another incredibly complex system, the human body. We can imagine a *Destination Human* initiative that aims to develop a highly accurate digital model of human health to monitor and predict the interactions between its physiological and pathological phenomena and human healthcare interventions.

To date, a *Virtual Human Twin* (VHT), a digital twin of the (entire) human body capable of predicting how the health status of any single individual may change due to internal pathophysiological processes or external interventions, does not exist. What we do have are *Digital Twins in Healthcare* (DTH), subject-specific predictive models designed to support a narrowly specific clinical decision. But it is easy to imagine how accumulating specialised DTHs, capturing fragments of causal knowledge, and digitally stored quantitative data, capturing empirical knowledge, can progressively evolve into a full-blown VHT.

To realise the VHT, it is necessary to simplify, standardise and accelerate the development of DTHs and the systematic collection of quantitative observational data on the health status of individuals over time, under the effect of different diseases, when exposed to a variety of healthcare interventions. The first necessary step in the definition of the VHT is thus an analysis of how DTHs are currently developed and what barriers are slowing down such development. After the discussion of the DTH definition, life cycle and barriers, the vision of the VHT will be presented along with the elements crucial to its realisation. The last section puts these elements in the perspective of the EDITH coordination and support action responsible for creating this roadmap.

4.2.2 Digital Twins in Healthcare: from generic to subject-specific

The management of human health in its broadest sense requires decision-makers to take **well-informed decisions that may affect the health status of single or groups of human beings** (hereinafter generically called *reference population*). Examples of this include clinicians making decisions on personalised therapeutic strategies for a patient; researchers making decisions on possible druggable targets to pursue in basic biomedical research; healthcare authority managers planning specific policies; biomedical companies seeking to refine, reduce and partially replace animal and human experimentation for the regulatory approval of new products; *etc.* This decision-making process usually involves the quantification of specific constructs that represent such health status called Outcomes, with selected metrics called **Quantity of Interest** (QoI), and then observing how such QoI develops in time, due to variations of internal (*e.g.*, body weight) or external conditions (*e.g.*, exposure to pollutants), or because of intentional interventions. The **Context of Use** (CoU) defines how the QoI informs a specific decision-making process relevant to human health and under which specific conditions such a process occurs.

QoIs are usually measured experimentally, either directly on human volunteers or patients or indirectly on surrogates such as animal or in vitro models. But these experiments pose a long list of practical, ethical, legal, and socioeconomic challenges and are primarily responsible for healthcare services' high costs and limited capacity. Thus, there is intense ongoing research on developing new technologies that can refine, reduce, and partially replace the need for experimental measurements to estimate the QoIs necessary to support decision-making within specific CoUs.

A DTH is a computer simulation that predicts (generating an output by a mapping from an input, as opposed to measuring experimentally) quantities of interest necessary to support decision-making within a specific context of use in healthcare.

Here, "computer simulation" refers to any software capable of predicting specific outputs given certain inputs. DTHs can be predominantly knowledge-driven predictive models built using existing knowledge, validated or hypothesised, about physics, chemistry, physiology and pathology or predominantly data-driven models built from large volumes of data using statistical modelling or artificial intelligence techniques, or any combination. In other industrial sectors, the term digital twin refers to a real-time computer simulation informed with sensor data. In the context of healthcare, this is not necessarily the case, although it might be. It should be stressed that while the DTHs can predict QoIs that are difficult or impossible to measure, they can do a lot more: they can predict how QoIs will evolve in time, how they will change depending on external actions, *etc.*

One of the most important features of digital twins is the accuracy with which they predict such quantities; thus, DTH can be divided into three broad categories:

- **Generic DTH**, for which the expected accuracy is that the predicted value is within the range of the values measured experimentally in the reference population;
- **Population-specific DTH**, for which the expected accuracy is that the predicted value is sufficiently close to some central property (typically mean or median) of the range of the values measured experimentally in the reference population;
- **Subject-specific DTH**, for which the expected accuracy is that the predicted value is sufficiently close to the value measured experimentally in each individual in the reference population.

By "sufficiently close", we mean that the predictive accuracy of the DTH is sufficient for its purpose as defined in the CoU. Thus, the same DTH can be sufficiently accurate for one CoU and insufficiently accurate for another.

Currently, most DTHs are designed to predict just one or a small number of QoIs with the necessary accuracy only in a narrowly defined reference population (*e.g.*, women over 55 with osteoporosis and no other conditions). This is because to develop DTHs, we need large volumes of detailed empirical observations and/or reliable mechanistic knowledge of the physiology and pathology/pathophysiology of the organs, tissues, and cells involved, as well as the mechanism of action of any intervention involved. Because of gaps in knowledge and data, the only way to manage this complexity today is to narrow the scope of the DTH, focusing on a particular process affecting a minimal portion of the human body, and to be used for narrowly defined CoUs. While these narrowly focused DTHs are extremely useful in specific cases, the time and cost required to develop DTHs with a broader scope and wider applicability are currently prohibitive. The Virtual Human Twin can provide such a framework, hence permits to derive a new generation of DTHs, capable of predicting any QoI necessary for any relevant CoU and reference population.

4.2.3 Digital Twins in Healthcare: the life cycle

The development of a DTH is a long and cumbersome process. It starts with the **identification of the clinical needs** expressed through epidemiological evidence that quantifies the limits of the current standard of care. HeartFlow is one of the first DTHs adopted in clinical practice[55], addressing a clear clinical need. Even though there is universal consensus among cardiologists that the best way to choose the most appropriate treatment for coronary stenosis is a Fractional Flow Reserve (FFR) measurement obtained through an invasive diagnostic test, only 20% of the UK patients with this condition are treated based on an FFR. HeartFlow provides a quantification of the FFR based on medical images. Another clinical need addressed by a DTH is osteoporosis, where the standard of care requires specialists to

[55] Rasoul H *et al*., Clin Med (Lond). 2021;21(2):90-95. https://doi.org/10.7861/clinmed.2020-0691 .

decide whether to treat a patient using DXA-aBMD as a predictor of hip fracture risk. With this risk predictor, around one-third of the patients are not treated; of those, around 50% will experience a hip fracture in the following five years. Considering that current treatment can reduce the incidence of hip fracture by around 50%, with a better risk predictor, up to 7.5% of all hip fractures (60,000 per year only in Europe) could be avoided. The Bologna Biomechanical CT-Hip (BBCT-Hip) DTH estimates fracture risk based on computer modelling and simulation on personalised patient data considering a wide range of fall scenarios[56].

The second step is **determining the causal relationship** between the QoI (hereinafter generically referred to as the model's outputs) and the parameters that control it (hereinafter generically referred to as the model's inputs). Living organisms are *entangled*, meaning that each internal state variable depends on several internal state variables. Of those, some variables have a greater effect on the QoI than others. Ideally, for each QoI, one should run a sensitivity analysis to analyse how the variation of any other possible QoI in the human body affects it. Since such systematic exploration is impossible, we use the available causal knowledge about the human body's physics, chemistry, physiology, and pathology to identify the minimum set of inputs that would allow a reasonably accurate prediction of the desired outputs. In some cases, the desired input cannot be measured on a patient-specific basis., In such cases we may build a statistical model for such a quantity that describes how it varies across the population of interest, possibly as a function of the inputs that can be measured for each patient. If this is not possible *e.g.* due to the lack of information, we may vary the desired input in the range of values observed in the reference population and study its effect on the QoIs.

When the available causal knowledge is insufficient to build a reliable predictor, data-driven modelling techniques can be used to identify the best possible predictor from all available inputs. But **training a predictor** requires a large collection of data, both in depth (data from many diverse patients are required to train the predictor) and breadth (as we do not know *a priori* which quantities govern the QoI, we need to explore as many as possible). Prospectively, creation of a DTH will in many cases require improving causal knowledge for example, by eliminating the competing mechanistic hypotheses or complementing existing hypotheses, in which pre-stages of the DTH may assist by guiding data acquisition.

The third step is the **model's implementation**. This is essentially a software development process that must be performed with the highest possible quality assurance efficiency. A key factor here is the availability of accurate input data to build benchmark problems used in the verification of the solvers. Another aspect is the definition of the model's execution environment. Depending on the nature of the data, there might be ethical-legal constraints imposing the data to be stored only at specific locations and under certain levels of cybersecurity; depending on the model implementation, there might be computational requirements that impose that the model executes only on specific computers. Given the expected complexity of a VHT as already of a DTH, software licences should be chosen to promote collaborative development and use in as early as possible development stages.

The fourth step is the development of all necessary **pre-processing and post-processing tools**. Pre-processing tools are those that extract the necessary inputs from the available data. For instance, we might need the volume of a tumour, which can be measured on a 3D MRI dataset, but only once the tumour is segmented in the images. The accuracy and the degree of automation of pre-processing tools are critical, and frequently excellent models are poorly informed by sub-optimal pre-processing tools. Post-processing tools are required when the model's output is not the QoI required to support the clinical decision-making process optimally. Optimal use of a DTH is strongly favoured by embedding it in a well calibrated pipeline of data acquisition modalities, pre-processing, DTH and post-processing.

The fifth and most important step is the **model's credibility assessment**. This vast and long process requires first data from tightly controlled experiments to conduct the verification, validation, and uncertainty quantification. There is also a need for data that quantifies the range of applicability of the model when used in clinical practice. Once the technical validation is completed, additional clinical validation might be required. This should be done independently from those who developed the DTH through the employment of prospective clinical studies data. However, in some cases, the regulator may accept studies on registry data (as far as publicly available) as evidence of clinical validity. In the

[56] Keaveny TM *et al*., Osteoporos Int. 2020;31(6):1025-1048. https://doi.org/10.1007/s00198-020-05384-2.

regulatory space, there is also an ongoing discussion on the possibility of certifying a DTH by allowing the regulators to conduct validation studies against publicly available experimental data.
The last step is the **provision of (clinical) access**. DTH can be made available to the end-users as software embedded in the medical imaging consoles, installable software, software-as-a-service, *etc.* This is also related to the business models to make the DTH widely available, which should include not-for-profit modalities (for example, clinical end-users funding the further development of specialised DTH by no-profit organisations).

4.2.4 Barriers to the development and adoption of Digital Twins in Healthcare

A good starting point to conduct this analysis is a similar one conducted by the In Silico World Consortium, which aimed to identify the barriers slowing down the adoption of In Silico Trials (IST). In Silico Trials are digital twins of cohorts of patients, which are used to assess the safety and efficacy of new medical products before their widespread use. This analysis identified seven barriers: lack of advanced models, lack of independent validation collections, no clear regulatory pathways, poorly informed stakeholders, poor scalability and efficiency, lack of trained workforce, and lack of accepted business models. Re-analysing these barriers in light of the VHT, we propose the following list of important barriers that need to be addressed for the VHT to realise its full potential in advancing human health.

1. Lack of high-quality data available in open access for the development and validation of DTHs
2. Lack of software components for the development of DTH available under an Open-Source license
3. Difficulties related to multiscale/multisystem models by available data at all space-time scales and for all pathophysiological processes.
4. Lack of Open-Source software libraries and standard operating procedures (SOPs) that simplify the deployment and the certification of the ICT infrastructures for the provision of clinical access, supporting distributed storage and execution models.
5. Lack of consolidated existing regulatory pathways with the EU and lack of harmonisation of a technical standard equivalent to the ASME VV.40:2018.
6. Lack of legal clarity and certainty for the clinical deployment of DHTs
7. Lack of well-established operating procedures and supporting data to demonstrate the efficacy and cost-effectiveness of DTHs.
8. Need for recruitment of medical technology experts by healthcare providers and their professional recognition as co-decision makers in the hospital DTHs investments.
9. Need for exploring strategies supporting a smooth transition from the pre-competitive to the competitive development of DTHs and developing value propositions to attract the investments of SMEs and large med techs.
10. Need for creating training and re-training programs on developing, assessing, and using DTHs.
11. Lack of well-informed stakeholders

Barrier 1: Lack of high-quality data available in open access for the development and validation of DTHs

As explained in section 4.2, the entire development cycle for a DTH involves access to high-quality data. First, for many QoI, there is a lack of consensus and standards of what high-quality data are and how they are produced, leading to dependency of data from the laboratory and even the handling person. The consequence can be missing reproducibility, conflicting interpretations and conflicting hypothesised mechanisms.
Second, since data is generally not accessible, each development team must invest significant time designing and conducting the experiments to produce the necessary data. Most of these data are then kept private, as they are perceived as the team's main competitive advantage over the others. Sharing in open access the experimental data generated to develop AND validate models is essential for the faster development of DTH. The development of Open Access, high-quality validation collections should be funded by funders and regulatory agencies, which could use them as independent validation evidence when comparing DTH targeting the same clinical problem.

Barrier 2: Lack of software components for the development of DTH available under Open-Source license

The first DTHs were artisanal software artefacts, where the developers implemented each required function from scratch. But as the field matures, this approach causes "the same wheel to be re-invented many times". This is also caused by the very low re-usability of DTH academic software and limited use of Open-Source licensing. One possible reason for this may be keeping competitive advantage over other academic teams, which plays an important role in attracting funding, often necessary for software maintenance. Another reason can be plans or an institutional strategy towards software transfer to companies or formation of start-ups, in which case closed software or transfer of exclusive user rights is usually strongly favoured by investors. In an ideal scenario, a DTH developer should focus on the core business of its model, reusing existing software for all other ancillary functions. To reach this situation, a balance needs to be found concerning open source software solutions facilitating collaborative work for DTH developers, and how this can be made compatible with functioning business models, which usually requires the attraction of investors. Ideally this includes a suitable balance for developers having contributed to a DTH software. Such solutions need to be embedded in SOPs enforced in all institutions that have contributed to a DTH software to avoid long-term delays due to conflicts between or inactivity at institutions. Both, functioning business models as participating the developers in profits in case of economisation, perhaps via standardised mechanisms implemented institutionally, should contribute to ensure high model and software quality standards at every level, as well as software administration and maintenance.

Barrier 3: Difficulties related to multilevel/multisystem models by available data at all space-time scales and for all pathophysiological processes.

Probably the most significant barrier is that the complexity of the effort required to develop a DTH increases exponentially. This concerns the aspect of the model creation as well of its unique description. The first generation of DTH focused on problems that were very much confined to space-time and physiological sub-systems. The accurate prediction of the FFR of a stenotic coronary artery can be obtained by modelling the phenomenon at a single space-time scale and considering only the physiology of the cardiovascular system (most parts of which can be lumped into some well calibrated multi-element Windkessel model, as most of the system dynamics is based on well-understood physical laws). Predicting the volumetric growth rate of a solid tumour as a function of chemotherapy is a much more challenging problem if the multiple space-time scales (whole tumour, cell-to-cell interaction, single cell system biology, molecular dynamics of binding affinity) are taken into account, which for accurate predictions aimed at within a VHT-derived DTH should be the case. Such a DTH will include how the tumour biology interacts with its environment , as well as how tumour cells penetrate, detach, are transported to other organs, and form metastases, requiring accounting for the pathophysiology of a good part of the human body. With each scale and each physiological sub-system added, the amount of data and knowledge required to build the model increases significantly as does the challenge of assessing its credibility.

A common solution to this problem is a "divide et impera" approach, where each scale, each physiological sub-system, is modelled independently, and complex multiscale/multisystem models are built as orchestrations of such "atomic" component models. Such an approach may be justified by existence of functioning submodules at each level. Some authors call these orchestrations *hyper-models* to stress that they are models of models.

With regard to model description in system biology, the reuse of models as components for more complex models has been addressed using standardised modelling languages (markup languages) and ontologies to ensure interoperability (*e.g.*, SBML, CellML[57]). However, this extremely elegant approach works well only for models where the mathematical representation and its numerical solution can be completely separated (as it is for algebraic ordinary differential equations). In field problems where partial differential equations must be used, the complexity of adopting this approach becomes considerable and has yet to find widespread adoption. A second approach, explored in some research projects, is to define the hyper-models as orchestrations of remote procedure calls, such as web services.

[57] https://www.cellml.org/

But here, the limiting factor is the complexity of developing, maintaining, and adopting the necessary orchestration libraries (*i.e.*, MUSCLE2, or VPH-HF v1 or v2). The simplest approach is to build the orchestration only in terms of data flow. Traditionally, this approach assumes that intermediate data objects are not useful and, thus, are not stored permanently (*i.e.*, Taverna Workflow Manager). But suppose one imagines a very large persistent dataspace; in that case, all DTHs could be reduced to atomic component models, as the data from other scales or sub-systems would be available as inputs. Of course, this would imply that the data are available in the dataspace at the space-time scale at which they were measured/predicted and homogenised/particularised at upper/lower scales.

Barrier 4: Lack of Open-Source software libraries and standard operating procedures (SOPs) that simplify the deployment and the certification of the ICT infrastructures for the provision of clinical access, supporting distributed storage and execution models.

A single DTH use needs to solve a single digital twin, whereas a single IST use may need to solve hundreds or thousands of digital twin models. Hence, while poor scalability and efficiency are significant challenges for advancing IST, for DTHs, the problem is more about providing *clinical access*. Each DTH solution needs to balance the need to keep sensitive data at prescribed locations and with the need to use appropriate computational resources usually unavailable within the hospital. Additionally, for the VHT to generate (clinical) impact, access to the computing infrastructure network needs to be streamlined and facilitated.

Barrier 5: Lack of consolidated regulatory pathways with the EU and lack of harmonisation of a technical standard equivalent to the ASME VV.40:2018.

DTHs are seen, from a regulatory point of view, as "Software as Medical Device" (SaMD) with predictive capabilities. FDA has recently clarified that also for those, the credibility can be assessed following the ASME VV-40:2018. IEC and ISO are working on a similar standard, which could be harmonised in the EU regulatory system. However, the VV-40 is recommended, even if not an EU-harmonised standard.

Barrier 6: Lack of legal clarity and certainty for the clinical deployment of DHTs

The use and deployment of DHT technology in a clinical setting requires stitching together the legitimate expectations of societal protection, the obligations of compliance with the different applications (and evolving) regulations and the ethical demands underlying the ongoing technological developments. Establishing a solid and uniform regulatory and legal framework will provide certainty and clarity to technology developers and providers and reassure the legal teams and investors. There is a need for a concrete and coherent environment to improve the development of DTHs by defining the legal criteria to be adopted to allocate risks and liabilities.

Barrier 7: Lack of well-established operating procedures and supporting data to demonstrate the added value, efficacy and cost-effectiveness of DTHs.

Paramount to the uptake of DHTs in clinical practice is the ability to demonstrate their added value compared to current standards, their efficacy and cost-effectiveness. For clinicians, we need to provide clear, conclusive evidence that using the DTH significantly improves the current standard of care (added value), and its effort is doable and justified (efficacy). For Payers (healthcare authorities, insurance companies, *etc.*), we need to demonstrate the cost-effectiveness of the DTH when compared to the standard of care. Currently, only anecdotal information is available on either aspect.

Barrier 8: Need for recruitment of medical technology experts by healthcare providers and their professional recognition as co-decision makers in the hospital DTHs investments.

Depending on the role of the DHT in the medicinal product life cycle (as a generator of digital evidence or as the product itself), different business models need to be developed. SaMD-DHTs could be considered health technologies that hospitals buy as instrumentations or services through competitive bids. But providing a clinical access model plays a significant role in deciding the commercial positioning. DTH developers can position themselves as Original Equipment Manufacturers (OEM) for established medical technology providers, as medical technology providers that sell instrumentation to hospitals, or as service providers with various business models linked to this. One apparent approach

would be running DTH technologies in clinics in analogy to MRIs, CTs *etc.* with specifically trained people consulting clinicians in applying the DTH technologies.

Barrier 9: Need for exploration of strategies supporting a smooth transition from the pre-competitive to the competitive development of DTHs and develop value propositions to attract the investments of SMEs and large med techs.

Key players are clear that, even though DTH models are often still at relatively low TRLs, the substantial current R&D investments into DTH science are insufficient to establish an economic sector. At the same time, current business models, licensing schemes and patentability are often not well adapted to sufficiently promote the progression of DTH development in academia towards commercial products. Instead, the need is for an economic value chain and ecosystem around DTH, with goods and service producers and consumers, prescribers, payers, supporting players such as certifiers, data or computational resource providers, and consolidated distribution channels. Implicit here is an entrepreneurial understanding of the pathways leading from scientific innovation to monetisation. Yet, it is not surprising that this understanding is difficult to achieve, because of several peculiarities of the DTH sector, given the scarce amount of experience that can be carried over from other sectors. Equally, it is impossible to provide a shared understanding of business mechanisms specific to DTH by analysis of existing business ventures alone because there simply are too few, and many are still regularly pivoting their models. In close concertation with the industry, value propositions for the VHT must be developed and a clear understanding of the balance between public investments/infrastructure and market/commercial dynamics. Developed value chains should properly balance the interests of the stakeholders of contributing

Barrier 10: Need for creating training and re-training programs on developing, assessing, and using DTHs.

We need experts in the companies developing the DTHs, in the notified bodies that CE-mark DTHs, in the HTA authorities that evaluate their cost-effectiveness, and in the healthcare providers to plan and steer the necessary technological investments toward this area. Linked to this last point is a delicate issue of where the decisional procurement power lies within healthcare providers.

Barrier 11: Lack of well-informed stakeholders

Adopting silico medicine solutions involves a long list of stakeholders, from the citizens/patients to policymakers. Virtually none of these non-technical stakeholders is well informed on the potentials and limitations of in silico methodologies. Various polls suggest opinions oscillating from the excessive expectations caused by the hype for AI to a principled rejection of the whole concept of predicting health. We must ensure all these stakeholders receive accurate, factual and adequately balanced information.

4.2.5 Vision for the virtual human twin

From the analysis of the life cycle of a DTH, and of the main barriers to its wider adoption, we can start to formulate some elements of the Vision of what the Virtual Human Twin could be. Other valuable elements come from a critical review of how the available knowledge on the human body is currently used in healthcare decision-making. Two significant issues stand out: not all knowledge is actionable, and too much knowledge is produced and used under severe reductionist idealisations. Also, a lot of data is collected cross-sectionally, losing the dynamics over time of the processes.

The simplest way to form knowledge on human health is to collect observational data on individuals at various points in their life span, in healthy and diseased status, and with/without the effect of specific healthcare interventions. But this is limited by several methodological, ethical, and legal difficulties: as a result, only a small fraction of this knowledge is *actionable*, *e.g.*, can be acted upon to solve practical problems such as healthcare decision-making. Tentative knowledge is considered actionable when its ***credibility*** is high enough to be helpful in solving that problem. But this is frequently an issue:

- Many available observational **datasets are qualitative or semi-quantitative**, even though the phenomenon of interest can, in principle, be expressed quantitatively. This makes the quantification of the credibility of the resulting tentative knowledge problematic.

- Much data is of **low quality or, worse, of unknown quality**. The latter case makes it impossible to estimate the credibility of the resulting tentative knowledge.
- Most **observational data is obtained from model organisms** such as rats, mice, fruit flies, worms, zebrafish, *etc.* But these represent empirical knowledge only for the organism they have been observed in. These organisms become models when we use them to infer knowledge about human health. Before considering the information, they produce as tentative knowledge, the *analogy* (functional similarity) between the animal model and the human target, must be demonstrated, which is rarely done (one example being Data Resource Center of the the NIH-funded SPARC project[58]).
- Another problem that we need to overcome are the **too many reductionist compromises** that impair medical sciences:
    - The first is the concept of the **average patient**. In the VHT, all knowledge should be referred to individuals at a given point in their life; population knowledge, where necessary, should be generated dynamically by averaging individual information. Same for probabilities over a time span, as required, for example, by epidemiology. This would enable a truly personalised medicine, where every bit of information (not only a subset, *i.e.*, the genome) is used to inform the medical decision.
    - The second is that of **scale separation**. For each physiological or pathological process, available knowledge stored in the VHT should cover from the atomistic scale to the whole organism scale, and sometimes, where the interaction with the environment is a paramount, even beyond. Homogenisation and particularisation models would capture the knowledge of how the data change depending on the space-time scale at which we observe them, providing a continuum representation of the knowledge over space-time. Finally, we could link the clinical signs at the organism and whole body level with the relevant microscopic events at the histological, single cell-, and molecular scales.
    - The third is that of **system separation**. Medical knowledge is organised into specialities[59], defined by the conventional separation of the body into 12 organ systems (cardiology, neurology, respiratory, *etc.*), by specific diseases (oncology, venereology, *etc.*), or by organisational roles (emergency, general). In the VHT, knowledge should be accumulated for all organ systems and their interactions, all disease processes, and every phase of the health care process. This opens to a truly holistic medicine still rooted in the principles of the scientific method and western medicine.
    - The fourth reductionist compromise is that of **age separation**. Medical knowledge tends to separate life span in paediatric (< 16yo), adult, and geriatric (>65yo). But there is solid evidence of *lifelong health*: what happened in our childhood influences adult and geriatric health. In the VHT, the entire clinical history of an individual, from birth to death, should be captured.

From these sparse considerations, we can attempt a first definition of vision:

The Virtual human twin is an integrated multi-level, -time and -discipline digital representation of the whole body enabling the comprehensive characterisation of the physiological and the pathological state in its heterogeneity, allowing patient-specific predictions for the prevention, prediction, screening, diagnosis and treatment of a disease, as well as the evaluation, optimisation, selection and personalisation of intervention options.

More practically, *the Virtual Human Twin is an ever-growing accumulation of all quantitative knowledge on how individual subjects' health status changes over time. All knowledge is digitally stored in the form of adequately annotated data and predictive models. Observational data capture empirical knowledge, whereas predictive models capture causal knowledge. Data and models must be annotated with enough information to assess their credibility.*

Who is going to realise this vision? As this challenge is beyond the capabilities of any single organisation or even community (*e.g.*, in silico medicine, wearables), we are convinced the VHT can only be realised by establishing and engaging the entire ecosystem. The notion of ecosystems captures the complex set of interlinkages among sectors and industries. The ecosystem encompasses all players operating along a value chain: the smallest start-ups and the largest companies, the research activities,

[58] https://doi.org/10.3389/fphys.2021.693735

[59] https://en.wikipedia.org/wiki/Medical_specialty#List_of_specialties_recognized_in_the_European_Union_and_European_Economic_Area

the services providers and suppliers, the users and all other stakeholders. There are several key elements related to the ecosystem that requires further elaboration, the first of which is the ***community of practice*** (CoP), which includes academic and industrial researchers, developers of DTH solutions, clinical users, industrial users, CROs, regulators, health authorities and payers, policymakers, *etc.* A stub for this community already exists: the In Silico World project has recently transferred to the VPH Institute an online community of practice called ISW_CoP, based on Slack, which hosts the conversations of over 600 experts on best practices for in silico medicine. The VHT community must define best practices to make the VHT vision a reality through consensus processes such as this road-mapping effort. While the definition of these best practices is a continuous process, some educated guesses can be already made, which help formulate the VHT initiative's mission.
To realise the VHT vision, the VHT CoP needs to articulate a mission formed by three additional key elements:

- The ***Infrastructure*** provides a concrete operational space within which the VHT knowledge is generated, stored, annotated, revised, integrated, exchanged, *etc.*
- The ***Standards*** capture the consensus on how things must be done to ensure the highest possible level of interoperability and reuse.
- ***Long-term sustainability*** defines how the ecosystem collaborates and evolves to fulfil this shared vision.

Thus, the VHT is a vision that can only be realised by following an ecosystem approach, including establishing a CoP, an infrastructure, a set of standards, and a tangible outlook on long-term sustainability.

4.3 The VHT ecosystem

4.3.1 The VHT community of practice

The stakeholder groups can be defined by looking at the typical value chain for medical technologies:

- **Researchers** who develop new knowledge and new methodologies and test them in pre-clinical and clinical settings;
- Teaching staff, who use Digital Twins during their lectures for demonstration purposes;
- **Innovators** who translate research results into potential solutions for clinical or industrial unmet needs;
- **OEMs** (Original Equipment Manufacturers), typically software developers that provide libraries and solvers used to implement the Digital Twins;
- **Business angels and investors** who support the creation of new companies that want to sell Digital Twins;
- **CROs** (Contract Research Organisations) that assist in the conduction of clinical studies for the validation of digital twins but also that can use digital twins to design and optimise clinical studies of new treatments;
- **Animal Research Related Organizations,** which provide valuable input for the creation and validation of the Digital Twin through the employment of animal experiments
- **Medical Device regulators** that provide marketing authorisation for digital twins that are used as clinical decision support systems, but also qualification for digital twins used as medical device development tools;
- **Drug/ATMP regulators** that provide qualification for digital twins used as drug/ATMP development tools;
- **Conformity assessment agencies** (*e.g.* Notified bodies) designated by an EU country to assess the conformity against relevant regulations and applicable standards of digital twins for healthcare or other medical products developed with digital twins before they are placed on the market.
- **HTA authorities** that evaluate the effectiveness, cost-effectiveness, and reimbursement level of new digital twins for healthcare or other medical products developed with digital twins;
- **Digital Twin vendors**, which can be categorised depending on their business model:
    - *Biomedical instrumentation sellers* that sell the digital twins as software embedded in their hardware;
    - *Medical device sellers* that sell digital twins as algorithms embedded in their medical devices or that complement them;
    - *Medical software sellers* that sell digital twins as stand-alone products or services;

    - *Broker sellers* that sell digital twins developed by third parties, usually as software as a service (SaaS);
- **Data brokers** that curate and resell use licenses for data collections usually generated by third parties;
- **GDPR officers**, including data controllers, data providers, DPOs, supervisory authorities, and their legal advisors, involved with the handling of sensitive data used to develop, validate, or use digital twins.
- **Buyers and Payers** that buy digital twins in healthcare technology to provide healthcare. This is a variegated galaxy of stakeholders linked to the national or regional healthcare provision model. It includes private and public healthcare providers, insurance, healthcare authorities, group buyers, *etc.*
- **Medical Product Developers** who use digital twins to design, optimise, or assess the safety and efficacy of new medical products, both in the pre-regulatory and regulatory phases, or train their employees using DT;
- **Healthcare policymakers** who may develop specific policies linked to the use of digital twins but also who may use digital twins to support policy making, usually with the mediation of experts;
- **Healthcare professionals** who's needs are an important driver and who may use digital twins for increasing their understanding of specific cases and co-morbidity, as clinical decision support systems or as a way to visualize (patho)physiology and therapeutic approaches to students and patients;
- **Hospitals and their IT departments** who will need to integrate digital twins in their daily operations;
- **Patients**, who's needs will be a major driver of the entire value generation chain and who will benefit by more targeted and personalised diagnosis, treatment, and health products pertaining to a better health state and better disease management.

For each of these stakeholders, the VHT is of interest but the value proposition can differ from one stakeholder category to another and even within a category depending on the exact role people have with their organisation or their interest. All stakeholders contribute to the development of the VHT, each with their own specific added value.

For researchers, the VHT represents the opportunity to continue focusing on their specific expertise while at the same time developing DTHs that encompass various organ systems, pathophysiology processes, *etc.* The slogan is: "holism by collaboration". In the early stage of development of the ecosystem, researchers will be the main contributors to data and models. Still, researchers-run repositories will play the lion's share even when the VHT becomes a more mature and industrially relevant infrastructure. DTHs will further serve in education and formation of medical students, doctors and health workers, and for pupils at school by teaching staff at multiple levels.

Innovators are expected to have a significant potential benefit. Taking to the market a DTH research prototype today takes between five and ten years; with the VHT, this time to market could be cut in half. They will contribute to the VHT by consolidating available resources into validated and accepted pathways for innovation, regulatory approval, *etc.*

OEM will use the VHT to better position their software layer and ultimately increase the revenues from the healthcare domain. They will provide solvers and libraries with a high level of maturity, technical and certification.

Business angels and investors will use the VHT as the primary source of information to guide their investments in the sector. Navigating the VHT, they will find the innovators with the most profitable products, develop new business opportunities as the sector structures and specialises, and guide start-ups and spin-offs toward the aspects of the business that follow their ideas and visions. They will give the VHT business wisdom and guide its development toward sustainable business models.

All Contract Research Organisations (CRO) will benefit from the VHT as DTHs reduce animal experimentation and refine and reduce human experimentation. But the most innovative CROs will be able to offer their customers a different level of service, where they support product development from discovery to post-marketing surveillance by maintaining data and models that support the entire development cycle of whole families of similar products. CROs will provide the VHT with the first level of commercial use and contribute to its long-term sustainability.

Medical Device regulators such as the FDA CDRH or conformity assessment agencies such as the European notified bodies will use the VHT to consolidate real-world data and in silico testing frameworks that will speed up but also drastically improve the efficacy of regulatory surveillance. They will contribute to the VHT by providing the highest levels of credibility through appropriate certification/qualification pathways.

Regulators of medicinal products such as FDA CDER and EMA will initially use the VHT to handle well-known regulatory challenges: These may for example be better-powered dose-response and safety

studies, placebo arms where the placebo is unethical, testing treatments for rare diseases, *etc.* DTHs may also provide non-invasive, more ethical alternatives to quantifying important biomarkers and refining human experimentation. But we hope they also use the VHT to slowly move their regulatory epistemology to less phenomenological paradigms. Drug regulators will contribute to the VHT with high complexity challenges around multi-organ, multi-system processes. For example, in vitro - in vivo drug effect extrapolations will be facilitated, as the VHT will naturally permit to take into account the multi-organ, multi-system drug effects that cannot be assessed in vitro.

Regulators of Advanced Therapeutic Medicinal Products such as CBER, and in general of all products that fit poorly into the classic device/drug separation, will finally have the opportunity, using the VHT, to develop a regulatory science more specific for these complex classes of medical products. They will provide the VHT with a drive to expand its limits to modelling all kinds of interventions, including gene and cell therapies, tissue engineering, *etc.*. Vice-versa, the development of the VHT will guide and drive development of data acquisition modalities, for example, because the VHT will permit the identification of the most informative missing parameters with regard to a certain objective.

Also, HTA authorities can harvest huge benefits from the VHT. In the short term, they will benefit from developing an in-silico-assisted HTA, where many questions can be first explored in silico. But in the long run, a fully integrated in silico development and quality assurance of biomedical products will enable a whole different HTA paradigm, where the transition from efficacy to effectiveness becomes a smoother, more continuous process, and the costs of post-marketing surveillance are drastically reduced (and thus its scope can be expanded without damaging innovation). The VHT will also make the pricing/reimbursement process more factual. HTA experts contribute to the VHT by expanding the initial focus on clinical use to a broader view of human health, which is also made of social, economic, and organisational determinants.

New companies that commercialise digital twins will see an explosion of marketing opportunities. They will monitor the VHT for pre-commercial solutions and guide the developers to their commercial exploitation according to their preferred business model. Biomedical instrumentation sellers will use the VHT as the innovation emporium, where to pick the next new thing to add to their product line. The VHT will make it easier to develop smarter medical devices that embed predictive capabilities and are supplemented by them. Medical software sellers and brokers will have an easy way to expand their portfolios, but also which solutions are worth investing in, in terms of credibility and popularity. Their role is key to the VHT, and they will drive to a more sustainable state.

The VHT will drive and consolidate the emerging business of health data brokers. Data brokers will use the VHT to find valuable collections or compose them by merging data collected by separate entities and assist them in transforming these resources into revenues. They will also contribute substantially to the long-term sustainability of the VHT.

The development, validation and use of DTHs currently pose several challenges for those legally responsible for treating sensitive data. The VHT will provide them with new solutions for these problematic activities regarding data protection and associated legal risks. Privacy experts will be essential in developing the ethical/legal framework within which the VHT will operate. They also advise the policymakers on any legislative intervention that might be necessary to make this aspect less critical than it is now.

Buyers and payers face significant challenges in handling disruptive innovations like DTHs. Leaving to med-tech giants to guide it will prevent exploitation that changes business models and may reduce costs while improving service. The VHT will make it easier to test and explore innovations in healthcare delivery models and build public procurements, even in the form of trans-European consortia. These experts will be critical in showing where the VHT can and does impact the European healthcare system.

The use of digital twins to develop and de-risk biomedical products, what we call In Silico Trials, has been and remains one of the biggest opportunities for medical companies involved with the VHT. The VHT will make the development, validation, and qualification of In Silico Trials methodologies easier, cheaper, and faster. Medical companies will drive the industrial development of the VHT.

Healthcare policymakers at all levels (regional, national, union) have a tremendous opportunity with the VHT to tap into a community of experts that can inform policy decisions with data, facts, and a broad spectrum of specialist expertise. Policymakers may become essential for developing the VHT, where legislative barriers or the lack of clear legislation may cause the VHT to develop more slowly and thus have a smaller impact than expected.

Finally and most importantly, the patient is in the centre of the entire endeavour. His needs will drive the entire development chain towards tremendous added value finally permitting a better and personalised diagnosis, treatment and management of disease and of aging-related functional declines. In addition, the VHT, and specifically DTHs will serve as an information source for the patient and patient initiatives.

4.3.2 The VHT infrastructure

What we describe here is a preliminary general vision of what, in the long term, the VHT infrastructure should be. This will inspire the proof of concept that the CSA EDITH is developing, but for obvious reasons, it will not be even close to this level of ambition. The **VHT infrastructure** should be developed around five strategic pillars considered essential for the global effort to drive forward **VHT** development.

1. Distributed/federated architecture
2. Governance
3. Openness
4. User Engagement
5. Industry collaborations and partnerships

Distributed/federated architecture

The heart of the VHT infrastructure will consist of a distributed/federated platform. The different services of the VHT platform will be semantically mapped to different types that occur within the context of the VHT. The platform will include centralised **core elements**, *i.e.*, the "hidden" elements required to run the platform. The second level will include the **platform and science-specific services**, *i.e.,* generic platform elements necessary to the end-users (*e.g.*, wiki, collaborative documents), scientific services like the repository, tools for running the workflows, *etc.* These might include services for semantic re-annotation or services to promote resources along the Credibility axis. The last level (**domain-specific services**) is end-user facing and includes all the federated services relevant to the VHT.

A future benefit of administratively treating **domain-specific** products and **services** separately from **Platform and Science specific services** is that it facilitates an easy onboarding path for more domain-specific tools and services and even tools from other scientific domains. The onboarding is further facilitated by the federated -more flexible- nature of the VHT platform. Component owners will enjoy full *autonomy* regarding the services and applications they provide. New services will be developed, mature, and be provided as federated services that will be part of the VHT platform, following integration/quality/interoperability requirements.

The VHT infrastructure needs to be "**as open as possible, as close as necessary**", *i.e.,* very accommodating of various existing components, formats, and protocols, but on the other hand, it needs to provide a unified and intuitive user experience that does not expose all the "sharp edges" of the underlying machinery and duct-tape. This follows the Robustness Principle (Postel's Law) made famous during the specification of the TCP protocol: "be conservative in what you do, be liberal in what you accept from others".

The implementation as a distributed/federated platform will be more **flexible** and **adaptable** to changing requirements and user needs. Different entities can choose which technologies and standards to use and evolve their systems independently. Moreover, its distributed nature will allow the platform to easily **scale** to handle increased traffic or user demands by adding more servers or nodes and to be deployed across **multiple locations**, allowing for a wider reach and better performance for users in different regions. Given the distributed nature of the storage services archiving data and the computing services elaborating on them, they need to be connected by high-speed geographic networks.

A federated/distributed infrastructure can also promote **interoperability** between different systems, allowing users to communicate and share data across different platforms, and fostering innovation. At the same time, it also improves the **performance** of the platform by reducing latency and increasing bandwidth. However, it may also require more coordination and **governance** to ensure interoperability and maintain quality.

Governance

Governance is essential since it defines the management structure, roles, and decision-making procedures. The **governance framework** will identify/establish the policies, governing roles and responsibilities (admin, provider, and user profiles) and decide the standards to adopt (section 6.2). It will also ensure the clarity of the business models and access policies (tiers, pricing policies, commercial agreements). It will build a detailed roadmap following the evolutionary ecosystem approach (section 7.3).

Openness

Openness in the VHT infrastructure will allow users and developers to share their work and collaborate with others, bringing together people with different backgrounds and expertise, leading to more diverse perspectives and insights, reducing, at the same time, duplication of effort and resources by allowing users and developers to build on each other's work. The research findings will be available to the entire CoP: the scientific community, the policymakers, and the stakeholders. An open VHT infrastructure will also facilitate the sharing of data and resources, encouraging, at the same time, uniformity of protocols and formats, standardised wherever possible and standardisable in hopefully all other cases (section 6.2). As an important element to guarantee openness, software shall be required to be licensed in a way that permits open access to the source code for those software entering the VHT infrastructure. This does not prevent transfer of rights for commercialisation from contributors to companies if necessary within business models.

User Engagement

A user-friendly and visually appealing platform will facilitate the CoP's interaction with the VHT infrastructure. The design will be user-centred, and user surveys will be used to gather insights that inform the design and functionality. The layout will be easy to navigate. The platform will have a dashboard that will guide the user through the different offerings of the VHT, allowing the users to customise their profiles and add the services and tools that are useful to them. Clear documentation will be available, not only for the main functions of the platform but also for the different services. The platform will combine tools and services from different sources, all valuable for developing the VHT. Users will use the platform also for collaboration and interaction with other users. The ecosystem will provide user incentives for sharing data, models, or other content (sections 6.5 and 7.3), while it will consider the feedback from the users, working to improve itself. Finally, user engagement should be measured! Analytics tools will be used to monitor user behaviour and identify improvement areas.

Industry Collaborations and Partnerships

As mentioned in the previous section, the industry can provide valuable guidance and feedback for designing and developing the VHT infrastructure. This collaboration can help ensure that the following steps focus on developing technologies and solutions that are relevant and useful to the real world while advancing the field of virtual human twins. By providing real-world context, the industry can help to better understand the challenges and opportunities associated with VHT. At the same time, the industry can provide expertise and resources (access to data, software tools, *etc.*). Their feedback will help develop more relevant and valuable solutions for external users while advancing the general field of VHT.

4.3.3 The VHT Standards

The VHT will heavily rely on standardisation to ensure the highest level of interoperability. Where possible, we will support **formal technical standards** developed by international standardisation bodies; where these are missing, we will use ***de facto* standards**. The Community of Practice will also produce standardisation by generating Standard Operating Procedures (SOPs) that codify the good practices and the rule of use of the VHT, as well as of requirements to software integrated in the VHT platform.

Lastly, the VHT will be part of a European ecosystem of information technology services such as the European Health Data Space. Here we commit to achieving the highest possible **integration and interoperability** with such services, including adopting specific standards that these services support.

This said, our view on the use of standards is pragmatic. It is essential to state that the goal is to ensure the broadest possible adoption for the VHT. So, when standards help this by simplifying access and ensuring high levels of interoperability, they are welcome; when the support of particular standards involves a significant overhead that complicates the adoption of the VHT, they will not be supported. Similarly, if the community is split relatively evenly between two *de facto* standards, we will support both, asking the promoters to work on translation tools that facilitate moving from one to the other.

4.3.4 The VHT Long-term sustainability

The interactions within the VHT ecosystem are driven by rules, policies, and standard operating procedures that regulate how the community of practice members exchange data, models, and services as they co-develop the VHT.

Given the complexity of the community of practice involved and the variety of value propositions they expect from the VHT, we **cannot imagine the ecosystem as rigidly fixed**. At the risk of oversimplifying, we can distinguish three phases, which presuppose the initial realisation, within VHT infrastructure, of a system of Distributed Ledger Technology (DLT)[60] allowing to permanent trace all types of assets exchanged on the DLT, also tracking their provenance and securing the findability, accessibility, semantic interoperability, and reusability of all activated resources.

In the early stages, the DLT infrastructure will host exclusively pre-competitive transactions and work on incentives based on a form of quality scoring (**Honour ledger**); in the second stage, pre-competitive and competitive transactions will coexist, and exchanges will be facilitated through the issuance by the DLT infrastructure, of digital tokens with no direct monetary value, but operating as the scaffold on which symbolic prices can emerge through supply and demand of all assets traded, included the DLT services (**Token ledger**); in the third and final stage, the ecosystem will mature and specialise: while some entities dealing mainly with pre-competitive transactions will continue to exist, a growing number of subjects will increasingly focus on competitive transactions in the form of business-to-business exchanges, with prices set in Euros and no-more in tokens (**Money marketplace**). This will be further elaborated in Section 7.3. The same section also discusses in more detail other sustainability elements in the private or public (or mixed) domain, such as the development of marketplace services, different business models and establishment of a public research infrastructure.

4.4 Conclusion

TBD

[60] El Ioini N & Pahl C. OTM 2018. Lecture Notes in Computer Science (2018); 11230. https://doi.org/10.1007/978-3-030-02671-4_16

5 Technology for the Virtual Human Twin

5.1 Introduction

From a technology perspective, the VHT can be thought off as a collection of resources that can seamlessly be interrogated, integrated, and executed, as well as the underlying fabric to facilitate such actions. The basic resources are data and models, which can be integrated into workflows. These three resources are stored and executed in the VHT infrastructure made up of storage, execution, and network resources, organised and integrated into a repository and simulation platform, relying on tailored software stacks.

5.2 Organisation of resources

5.2.1 Multidimensional space as an organisational paradigm

The elements discussed in Chapter 4 provide a general outline for the blueprint of the Virtual Human Twin. This section provides a high-level description of the essential characteristics of the VHT. Every resource that forms the VHT is extensively annotated, and this metadata can be used for advanced searches and other complex operations. However, it is crucial to identify an organisational paradigm, which helps develop more coherent user interfaces, make it easier to access base usage, *etc.* This paradigm is to represent all VHT resources in a multidimensional space[61].

5.2.2 Definitions

The Data-Information-Knowledge-Wisdom (DIKW) pyramid, see Figure 1, is a conventional representation of how data become decisions.

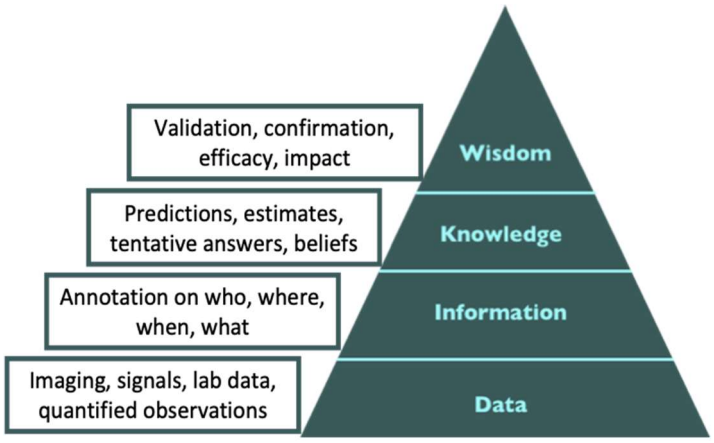

Figure 1: the DIKW pyramid

The VHT contains data, information, knowledge and, indirectly, wisdom. The VHT will allow uploading of datasets in digital format, as far as:

- The data refer to quantitative measurements;
- The data refer to human pathophysiology;
- The data are separated per each individual.

The only reason to relax these limitations is when no data of that kind exist. In those cases, we can tolerate that a dataset is uploaded that uses semi-quantitative scales (*e.g.*, pain scales), is obtained from animals, or is the average of measurements made on multiple individuals. But these datasets are particular and they need to be annotated as such to make sure the VHT users are informed.

[61] Viceconti *et al*. (2023); arXiv:2304.06678v1. https://doi.org/10.48550/arXiv.2304.06678

The VHT will not allow uploading data that are not annotated with the minimum set of metadata. This minimum set depends on the data type, but generally, it should answer the questions of what, who, where and when. So, this means that the VHT will allow only the upload of information, not data.

Here we use the term knowledge in the Bayesian statistical meaning of belief (prior probability), where its truth content is expressed in terms of probability. So, knowledge is not true or false, as in frequentist statistics, but more or less likely to be true, depending on the evidence of credibility that has been produced. In the VHT, knowledge is captured into quantitative predictive models. Wisdom is achieved when there is sufficient evidence that some knowledge is credible enough to support a specific decision-making process.

In the previous section, we identified the following resources as constitutive of the VHT:

- Data objects;
- Annotation Services;
- Model objects;
- Workflow objects;
- Execution, storage and networking Services.

The central resource is the data object. A model object is a relation between some input data objects and some output data objects. A workflow object is an orchestration of model objects executing over a set of input data objects. Data objects are annotated with their pose in the VHT multidimensional space; model objects and workflow objects will assume a pose that is a function of their input and output data objects. Below, these concepts are explored further.

5.2.3 The data object

Each VHT data object is a digital dataset, stored and annotated according to some basic rules. The dataset must contain quantitative information on human pathophysiology, whether measured or predicted. It must be stored and curated according to the FAIR principles to be findable, accessible (possibly through authentication and authorisation), interoperable and reusable. The dataset must be annotated with sufficient metadata, which includes information on the data object type and its position in the data space. The Data Object Type (DOT) is a unique identifier associated with enough information to decide if and to what extent that data object is suitable input for a DTH model. This includes information on the dataset regarding its semantics (what the data mean), its syntax (in which standardised, interoperable formats the dataset is accessible), and its accessibility (how the dataset can be accessed). Eventually, DOTs will be selected from a list of standardised types, possibly organised in a well-structured taxonomy or ontology. The list of supported DOTs might start as a folksonomy, a user-generated way of organising content, which is periodically scrutinised and consolidated into proper VHT ontologies.

In computer vision and robotics, the pose of an object is the combination of the object's position and orientation. Pose estimation determines a detected object's pose relative to some coordinate system. This information can then be used, for example, to allow a robot to manipulate an object or to avoid moving into the object. The Data Object Pose (DOP)[62] includes all information to define the position of the data object in the VHT n-dimensional reference system and the scale information, such as the grain and range[63] of the dataset[64]. The grain is defined as the larger of the minimum distance (or time span) that can be distinguished by the instrumentation or as the characteristic distance (or time span) of variation of the smallest (or fastest) feature of interest measured using this instrumentation. The extent is defined as the smaller of the maximum distance (or time span) that the same instrumentation can

[62] In computer vision and robotics, the pose of an object is the combination of the object's position and orientation. Pose estimation is the determination of a detected object's pose relative to some coordinate system. This information can then be used, for example, to allow a robot to manipulate an object or to avoid moving into the object.
[63] Grain is defined as the larger of the minimum distance (or time span) that can be distinguished by the instrumentation, or as the characteristic distance (or time span) of variation of the smallest (or fastest) feature of interest measured using this instrumentation. Extent is defined as the smaller of the maximum distance (or time span) that can be measured by the same instrumentation, as the characteristic distance (or time span) of variation of the largest (or slowest) feature of interest measured using this instrumentation.
[64] Bhattacharya *et al*., 2021

measure as the characteristic distance (or time span) of variation of the largest (or slowest) feature of interest measured using this instrumentation.

The six dimensions of the data space are represented in Figure 2 and defined in the following sections. The concept of grain and range as scale representation applies well to datasets that define the variation of a quantity in space and time. But since we assume by convention, as described above, that also scalar values are associated with a point in the 6D reference system of the VHT, in that case, the grain represents the least significant digit of the measurement/prediction (reproducibility of the measurement, uncertainty of the prediction). In contrast, the range could represent the uncertainty of positioning in space and time for that scalar quantity.

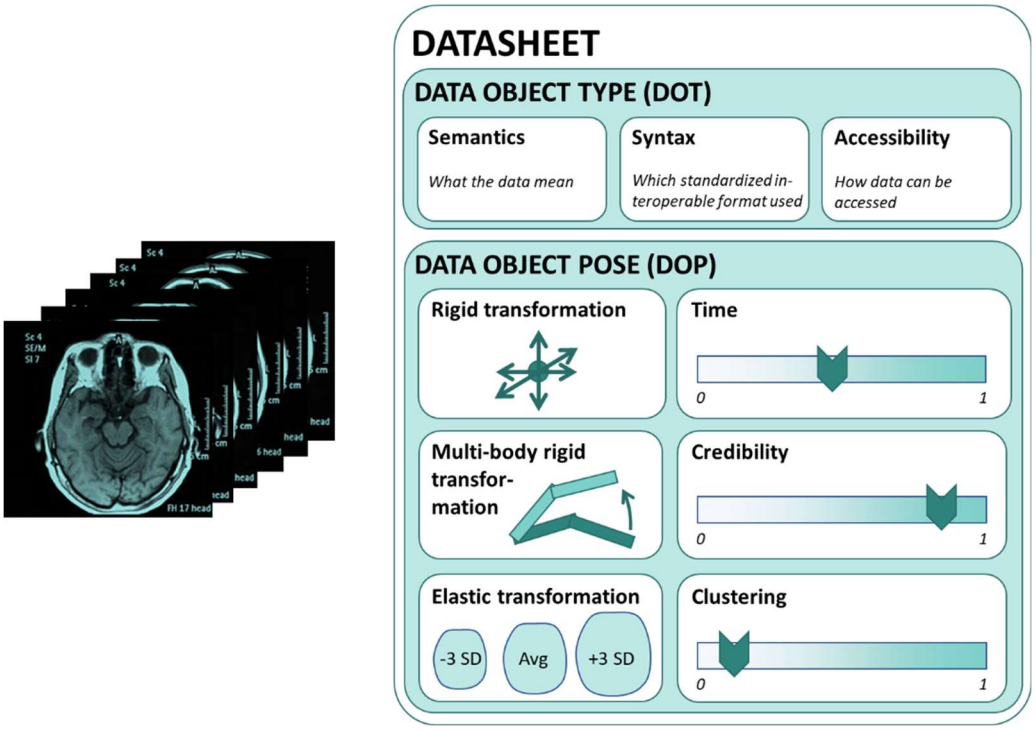

Figure 2: A graphic representation of the data sheet describing the relevant information of a data object

Whenever a new DOT is added, it should also be provided with the transformation functions required to calculate the DOP for each data object with that DOT. See Figure 2 for a schematic overview of the DOT/DOP and below for a more detailed description of the data space and its six dimensions.

A general data object template

Depending on how the VHT will develop, there might be some for developing a general data object template, of which every possible data object type considered in the VHT is a specialisation. Some related work was done as part of the development of the so-called Multimod Application Framework, a rapid application development library for biomedical software applications[65]. There such general data object template is called a Virtual Medical Entity and defined as a semantically annotated, time-varying posed, 3D bounded, time-varying multi-scalar field.

The data object pose

We provide an organisational paradigm by assigning a pose in the multidimensional space to data objects. We have identified, so far, six dimensions for the data space; others might be added as the VHT is being developed: Three spatial coordinates, time, clustering, and credibility. All this information combined defines the DOP.

[65] Viceconti M et al., Proc Eighth Interntl Conf Inform Visual (2004); 15-20, https://www/doi.org/10.1109/IV.2004.1320119 .

Space

The space is that of the human body as represented in anatomy. But where descriptive anatomy traditionally provides a qualitative, descriptive (semantic) representation of the spatial organisation of the body, we need a quantitative, universal representation of the anatomical space. This anatomical space is the average body geometry of all human beings, the average human body. We call this the VHT anatomical template.

Data objects can be defined over 0, 1, 2, or 3 spatial dimensions. For example, the systolic blood pressure of a subject is a 0D data object, how the blood flow velocity varies along the length of an artery is a 1D data object, the distribution of temperature over a region of the skin is a 2D data object, the distribution of bone mineral density in a bone is a 3D data object.

Each data object (except 0D objects) represents the spatial variation of its values using a reference system (implicit or explicit). So, in a 3D data object, the value corresponding to the coordinates (0, 0, 0) places such value at the origin of this reference system. In addition, each data object is referred to a specific individual and their anthropometry. But to simplify automatic annotation, clustering, and other similar operations, it is convenient that each data object is mapped to a conventional anatomical space by posing it with respect to the VHT anatomical template (which provides the DOP).

Using a conventional space requires each data object type to be provided with one or more annotation functions to transform it into the conventional spatial reference system. 3D objects can be easily posed in the anatomical space; the bone mineral density distribution of a patient's femur can be posed in the VHT Space region corresponding to the femur of the anatomical template. With some caution, 2D and 1D objects can also be posed with respect to the anatomical template. However, 0D objects do not have an anatomical location. But because all data objects in the VHT must have one, all 0D data objects are, by convention, mapped on a 3D point located in a conventional point in the anatomical space. So, for example, the systolic blood pressure value could be posed at the centre of the heart region in the anatomical template or in the arm region where the sphygmomanometer was applied.

For spatially localised data objects, it is necessary to define appropriate transformations to calculate the pose of the data object in the conventional anatomical space. Spatially localised data objects usually have a volume, and thus they occupy a region of the anatomical space, usually represented with the bounding box. As such they do not only have a position in space but also an orientation. Moreover, the spatial transformations are also necessary to produce the average data objects to populate the clustering axis. Because we want to preserve the integrity of the original data objects, and because applying a spatial transformation is usually not computationally expensive, it is convenient to keep the data object in its original spatial localisation and store the transformation parameters that pose it in the conventional anatomical space as metadata, that are applied on the fly to the data when required.

The minimum set of transformation parameters should include the roto-translation matrix that rigidly transforms from the reference system of a spatially localised data object to a pose that is anatomically appropriate in the reference system of the VHT anatomical template. Anatomically appropriate means centred and aligned. So, for example, the 3D surface of a femur would be transformed to have its centroid coincident and its principal inertia axes aligned with those of the femur of the VHT anatomical template. Of course, that subject's femur would generally be smaller or bigger than the average femur.

Users can also add the transformation parameters for more complex spatial transformation, such as an affine transformation that scales the data object so that its bounding box matches that of the template organ or fully elastic registrations that match the geometry of the data object to that of the corresponding template. Elastic registrations are used to generate average datasets for specific population clusters.

Among the essential metadata for each data object, one must include the spatial range and grain, which facilitates the definition of the spatial scale in which the data object is defined.

The anatomical mapping of all data objects to a conventional anatomical space poses some challenges. For example, how to handle datasets that refer to multiple anatomical locations (*e.g.*, recordings of a multi-lead electrocardiogram). In such cases, one could position the dataset in correspondence with the heart centre or the chest region's centre. Or, if the anatomical location of each lead is available, one could decompose the dataset into multiple data objects, one for each channel, and place them at the anatomical location of their lead.

Time

The time axis may require three different representations depending on the use case. The first is an absolute time axis where each data object is placed at a specific date/time coordinate (for example, using the W3C date and time formats specifications). However, in some cases, it is more useful to represent time as the age, the relative time passed since birth. For this purpose, we need to add a birth date, actual or estimated, to the metadata. Last, in other use cases, data objects could be organised along a normalised time axis that spans from zero (birth) to one (death). For this, we also need in the metadata a death date, actual or estimated. Since these three representations can be easily generated on the fly when the necessary metadata is available, we recommend storing the date/time at collection, birth date and death date, leaving it to the user interface to choose the most convenient representation depending on the use case.

As for space, time-varying datasets will range between two date/time points. However, storing the start date/time and duration might be convenient instead. Since the time of birth is not generally available information, we will assume that all subjects born on a given day were born at 12:00 (noon) because sixty per cent of babies are born during the day, between 6 A.M. and 6 P.M.

Among the essential metadata for each data object, one must include the temporal range and the grain of the data object, which makes it possible to define the time scale in which the data object is defined.

Clustering

The human pathophysiology varies widely between individuals. On the one hand, the VHT should contain quantitative knowledge about the pathophysiology of many individuals; on the other, in several situations, such knowledge needs to be represented (and in some cases is provided) only as an average of a cluster of individuals. Each data object type must include an averaging function that enables clustering among its annotation services. For data objects defined in space, this is typically an elastic registration function; for time-varying objects, it might involve a synchronisation function, and for data objects not defined in space-time, these are averaging functions in the statistical sense.

When added to the VHT, each data object is placed at clustering $k = 0$ (no clustering). To ensure irreversible anonymisation, the metadata includes a unique data object ID and a unique PatientID, not associated with the individual identity. Where necessary, a LocalPatientID can be used to support pseudo-anonymisation schemes.

All data objects are automatically added to one default cluster: homo sapiens ($k = 1$). Specific research projects may calculate other sets stored with enough metadata to inform the number of groups and the criteria used for clustering. This means that on the Clustering axis, there might be in the same coordinate multiple data objects for the same DOT type, each obtained with different clustering criteria. So, for example, under $k = 0.5$, we could have a male-female, healthy-diseased, or a clustering above or below 55 years of age. How do we map clustering rules with the same k value? One possibility is using the same interface used to handle multiscale data.

Suppose we navigate to the cell of CT scans of the femoral bone. In that case, we will find clinical CT in which the whole femur is depicted, microCT in which only a small tissue biopsy is depicted, and nanoCT in which only a few trabeculae are depicted. The user interface will have to represent the fact that in that cell there are data objects defined on a different space-time scale and provide a way to navigate this additional organisational principle. Similarly, on the Clustering axis we will see that for a certain k value there are multiple cells, each associated to a clustering rule with the same k value.

Credibility

When a new data object is added, it is placed at the lowest level of credibility (non-qualified data). The data owner can submit a data object to the credibility transformation function. The higher the credibility of a data object, the higher its value. Depending on the level of credibility that the owner is requesting, the application must be informed by a smaller or greater amount of information that captures the provenance, the quality, the metrological properties (or computational credibility properties if the data are computed), and the certifications of the instrumentation/software used. For high levels of credibility, the request might be evaluated by a panel of experts, possibly in coordination with regulatory agencies.

The disease state

Considering the main aim of the VHT is healthcare, one could wonder why there is no disease axis to the multidimensional space we use as an organisational paradigm. We can represent the human body as a closed system in a state defined at each instant by the values assumed by its state variables. Diagnosis, the identification of a disease, is not only characterised by symptoms (unusual combinations of values for some state variables) but also by some causal relationships. Indeed, in medicine the term syndrome indicates clusters of symptoms that might be associated with different diseases (differential diagnosis) or lack any causal explanation. But being healthy (or diseased) is also a subjective construct of the patient. So, the concept of disease is too complex to be used as an organisational paradigm. Still, VHT users will be able to annotate their data objects with disease-describing metadata using clinical terminology dictionaries such as SNOMED-CT[66], ICD[67], or full ontologies such as EBI's DOID[68].

An example

We use the hypothetical generation of the VHT Anatomical Template to illustrate how these six dimensions are defined. Let us imagine having a large collection of 3D body scans of humans of all ages, genders, *etc.* In theory, all scans were taken with the subject in the same pose (standing with the feet slightly apart, arms along the sides with the palms forward).

Each dataset is expressed with respect to an implicit reference system specific to the type of scanner used. We position the data object on the time axis in correspondence to the date/time of the scan, but we also add the subject birthdate, so we can represent the data objects also with the age axis.

Assuming the scans were all performed with fully certified 3D scanners, we place all datasets at value 1 on the credibility axis (which ranges from 0 for non-qualified data to 1 for fully certified measured data). Since each dataset refers to an individual, we place all of them at 0 on the clustering axis (as defined above).

If we now select all datasets for individuals of a certain age, we can perform some spatial normalisations. The first normalisation operation assumes the body is a rigid object. We define an anatomical reference system (*e.g.*, origin in the projection of the centre of mass on the floor, X oriented from posterior to anterior, Y from medial to lateral, and Z from feet to head) and calculate for each dataset the rigid transformation so that they are all aligned to the anatomical reference system. The second normalisation operation assumes the body is a kinematic chain, *e.g.*, a set of rigid bodies articulated through idealised joints. We define in the anatomical reference system an ideal body posture. Then we calculate the multi-body rigid transformation for each dataset that aligns each scan to this ideal body posture. The third and last spatial normalisation assumes the body is an elastic object. We use statistical atlas techniques[69] to calculate for each time point the average body shape and then calculate the transformation of each dataset to this average body shape. The vector of average body shapes at different ages is the VHT anatomical template. Each new VHT data object must be posed to this anatomical template.

[66] SNOMED-CT
[67] ICD
[68] EBI's DOID
[69] Sotiras *et al*., 2013

5.2.4 The model object

For consistency of the user interface, we should consider defining also for the model objects a Model Object Type (MOT) and Model Object Pose (MOP). The MOT must contain the list of DOTs from its input and output set. It must also contain all the metadata to ensure traceability, such as version, author, *etc.*

Location in the anatomical space, Age, and clustering does not apply to models. Model objects should be organised in the VHT space only along the Credibility Axis. However, if consistency with the data objects is considered useful, the MOP could be automatically calculated as:

1. Average of all poses of all inputs and outputs;
2. Average of all poses of all outputs;
3. A function in the multidimensional space that links some data objects (inputs) to others (outputs).

In VHT, model objects are defined as data space crawlers. A VHT model requires a finite number of inputs, described in terms of DOTs and DOPs, and produces, upon successful execution, a certain number of outputs, also described in terms of DOTs and DOPs. When a model is active, every time a new data object with the necessary DOT is added to the data space, the VHT model should be automatically executed. Its outputs are also added to the data space in the appropriate DOP. This is why they are defined as data space crawlers: we can imagine model objects like little insects that crawl the honeycomb of data objects, "eat" some data objects from certain honeycomb cells and "lay" some new data objects in other cells. Thus, every time we add to the VHT a group of data objects that constitute a valid input for a model object, the dataspace will automatically be enriched with new predicted data. This implies that VHT models must execute in batch mode. However, human interaction is still possible using the Mechanical Turk paradigm[70].

However, such "eager" execution model, while theoretically fascinating, would pose in practice a lot of challenges. A possible solution could be this: when a new valid input is added to the VHT all models that can use that input are automatically executed. But this only means that these simulation jobs are added to a queue, where they will stay indefinitely, until someone binds one job to two information sets. The first defines the computational resources that can be used to run that simulation. This means a specific simulation cluster, but also a valid account in it to which sufficient computational resources to run the simulation are associated. The second defines if the results of the simulation should be deployed first in the sandbox of the user requesting the execution (so they can inspect the outputs before publishing them) or publish them directly in the VHT.

There are two other important technical aspects that need to be addressed: remote execution and orchestration. The VHT will run on a single computer cluster with some storage in the simplest scenario. All data objects are stored in this storage, and all model objects execute on the computer cluster. But as soon as we imagine more complex architectures, we might have a situation where the storage that contains the data objects, and the computer that executes the model objects, are not co-located. To ensure maximum flexibility, we can imagine a scenario where both data objects and model objects are portable, the first using data replication services and the second using container architectures. This would allow the creation of a rule-based system that decides case by case if it is better to move the data or the models. The second issue is model orchestration, which is addressed by Workflow Objects.

Workflow objects

Complex phenomena are more easily modelled as orchestration of multiple models, each capturing a particular aspect of the knowledge available on that phenomenon. Typically, orchestrations define the data flow (how data is passed from one model to the other) and control flow (in which order and under which conditions the models execute).

[70] Yetisgen-Yildiz *et al.*, 2010

An orchestration is defined as strongly coupled when the models involved need to repeatedly exchange a very large amount of data. These models can be orchestrated using one of several available specialised libraries[71]. For efficiency requirements, this special class of orchestration is better exposed as a single monolithic executable, which internally invokes various codes that exchange data via memory.

In all other cases orchestrations can be represented with workflows. But workflows do not only contain control flow and data flow instructions. They can define data selection algorithms (which of the data objects available for the input types are to be used in the simulation), pre-processing and post-processing algorithms, *etc.* In a way, a workflow can be constructed (for example using knowledge graphs) to represent a use case, where selected data and processed with selected models and services to provide the necessary output. This could be very useful, in all cases where a decision does not depend only on data or models, but also for what they are used for enabling increased interoperability among models and better reproducibility across research teams.

If the control flow is a Directed Acyclic Graph (one model executing after the other), the workflow may be left implicit in the data structure, thanks to the VHT automatic execution.

5.2.5 The credibility axis in the Virtual Human Twin

The VHT contains two types of resources: information in the form of adequately annotated digital data and knowledge in the form of predictive models. When these resources are added to the VHT, they always start with zero credibility. As evidence of accuracy is added, their credibility can be increased according to rules defined by the community of practice for each type of resource.

The VHT will contain almost exclusively quantitative information obtained from humans and provided separately for each individual. For the few datasets that are semi-quantitative, are not obtained in humans, or are provided only in terms of population average, the credibility problem becomes extremely complex and articulated. In the first instance, our advice is to simply not provide mechanisms for credibility upgrade. In other words, all datasets that violate these constraints can only stay at credibility zero.

Information can be of two kinds: statements (*e.g.*, surname = Smith) and measurements. For statements, accuracy is a binary concept: the surname is Smith or is not. In some contexts, the accuracy of statements is called correctness. The accuracy of measurement is defined in a metrological sense; it is expressed by a pair of values, a measure of Trueness and one of Precision. Trueness expresses the systematic errors affecting the quantification; Precision expresses the casual errors.

The credibility of knowledge, when this is captured in predictive models, is expressed in terms of predictive accuracy against controlled experiments. But it also requires the decomposition of the predictive error in numerical, aleatoric and epistemic components through the process known as verification, validation and uncertainty quantification. However, in the most common case where numerical errors are negligible compared to the others, validation can estimate Trueness and uncertainty quantification Precision.

Thus, a more helpful categorisation for the Credibility axis could be if we divide data into Categoric and Quantitative. For categorical data, Credibility can be either 0 or 1, depending on if they have been checked or not. For quantitative data, Credibility is expressed in terms of Trueness and Precision, whether the value is measured or predicted. But because the concept of credibility is not independent of the use we do of the data, we recommend that even for quantitative data, the Credibility axis admits only a few conventional values, such as:

- 0.00 – no credibility;
- 0.25 – Partial evidence of credibility provided;
- 0.50 – Complete evidence of credibility is provided;
- 0.75 – Evidence of credibility is considered sufficient by the Community of Practice;
- 1.00 – Credibility is certified by a third party.

[71] Borgdorff *et al.*, 2014

In metrology, the validity of a measurement chain (and thus of all the information produced with it) is qualified by first measuring the same specimens with a measurement chain known to be at least one order of magnitude more accurate than the one we are testing (reference chain). We then compare the series of measurements done with the chain under testing with those obtained with the reference chain. This provides the accuracy of the information produced with that measurement chain. We then repeatedly measure the same specimen: the repeated measurement's variance provides the information's precision. Of course, accuracy and precision can vary depending on the measured value. Thus, a metrological campaign is conducted over various specimens where the measured value ranges between a minimum and maximum, which are considered the interval of validity of that measurement chain; accuracy and precision are provided as the average values over these repeated assessments. Thus, the credibility of quantitative information is fully characterised by identifying which measurement chain was used to generate it and the average precision and accuracy of this measurement chain over its validity interval.

In the VHT, knowledge is primarily contained in predictive models. Assuming that the predictive error is due to the sum of the numerical, aleatoric, and epistemic errors, we can also define the concepts of trueness and precision for models as per the information resources. Not all models are equation-based or use numerical methods to solve equations and for those models the numerical error is zero. Otherwise, it must be quantified. The goal is to demonstrate that the numerical error is negligible if compared to the sum of the other two such that its effect can be neglected. Thus, the most helpful information is the upper boundary of the numerical error. Once this is done, we can use experimental validation to quantify the accuracy and uncertainty quantification to quantify the precision. Same as for information, knowledge should be annotated with some limits of validity. These can be of epistemic origin (the knowledge used to build the model carries some limits of validity). Still, we think it is more helpful to indicate the limit of validity as the range of input values explored in the validation experiments to provide a basis for assessing the model's applicability in the ASME VV-40:2018 sense.

Traditionally uncertainty quantification aims to explore how the aleatoric uncertainty affecting the inputs propagates into the outputs. However, if we plan to use uncertainty quantification as a measure of precision, we should also include whatever residual epistemic uncertainty affects our model in the uncertainty quantification. The most concrete example is to include the uncertainty due to using some population averages as inputs instead of subject-specific values. But it also applies to environmental factors, *etc.*

5.3 Data

5.3.1 Data generation

Data generation originates in most **advanced omics, smart sensors (wearables and implantables), nanomedicine, advanced imaging, body-on-chip** and other **enabling technologies** to source the disease and organ models. Advanced nanomedicine techniques will further research in personalised healthcare with particular focus on biomaterials for regenerative medicine, targeted nanotherapeutics, and molecular imaging, with toxicity robustness and safety and lack of immunogenicity. In the context of the VHT, providing future advanced tools and infrastructure to accelerate the generation of new deep data that can further enhance our knowledge of the means to support prevention and personalised treatment of diseases, remains a very important endeavour. These can include, but are not limited to, engineering advances in energy-efficient sensing, computing, communication and cloud/fog technology together with adequate modelling and computational approaches such as Artificial Intelligence to generate and exploit big and deep data.

Overall, the VHT aims for an inclusive approach, to welcome all human data types generated from a diverse source of methodologies. We envision three broad categories of data in the VHT in terms of the source:

1) **Clinical Grade Data**: This encompasses data that is already being used as part of the healthcare and clinical practice. This data type is expected to be generated by clinical grade technologies and their validity is ensured by the healthcare system. Hence they do not require validation or justification to be

included in any modelling efforts. Hence, the quality of the data, data format, interoperability should be addressed in the VHT but the data type validity is not questioned. Examples are clinical biochemistry tests, MRI and CT scans, Electronic Health Records, vital signs (blood pressure, ECG, core body temperature). clinical genetic tests, key biomarkers specific to some medical conditions and to disease progression/regression, *etc.*

2) **Research Grade Data**: This category encompasses sufficiently advanced research grade technologies that are accepted in research practice across the globe. These technologies are expected to satisfy research community standards. Examples of such technologies are transcriptomics, proteomics, metabolomics, wearables with a combination of vital signs (heart rate, body temperature, SpO2, etc), activity tracking, emerging dynamic fingerprints of biomarkers in human biofluids detected by multimodal sensors in real time), implantable technologies and organ-on-chip technology (at long term, emerging in a body-on-chip technology). Another important category of data consists in exposome date (such as food, air, water but also other factors resulting from lifestyle and associated stress). Data types in this data category require evaluation for their fit for purpose and raise important challenges in terms of very heterogenous time-scale, format and interoperability.

3) **Data generated or transformed by models**: This category includes the results of simulations that are either transforming or generating data and creating a new data type. This category of data needs to be evaluated as their utility depends on their scope, sensitivity and accuracy which are affected by the model's input data types and model's precision and accuracy. Examples of these data types would be internal cellular or physiological parameters that could not be measured in-vivo, in-silico trial data, simulation results of a personalised DTH models, synthetic data generated with AI methodologies, *etc.*

Within the VHT Repository (see Section 5.6.1) we aim to link with existing data repositories required for the models of the VHT. Such data repository must be dynamically updatable and structured according to the proposed categories; data will be collected and stored in appropriate ISO 20691 compliant standard formats in virtual platforms from which they have selectively accessible via streaming.

Please note that the term "data" of the Repository is used for the repository as an alternative for "Research Object". A Research Object (RO) can be used to describe data, models, notebooks, workflows, services, software or even entire simulations (via Platform).

5.3.2 Platform access and Data management

Here we introduce potential Data Access, User Profiles and Roles within the VHT infrastructure. Our vision in designing VHT platform access and data management is to have the first blueprints of an *adaptable*, *secure*, and *efficient* system, equipped to answer the dynamics of healthcare across EU borders, utilising existing data, model and state-of-the-art computational resources to leverage the revolutionary potential of DTH.

While designing user profiles and user roles and their capabilities within the VHT platform we pay special attention to :

- **Efficient Allocation of Resources:** Understanding the specific needs and usage patterns of different *roles* allows us to allocate resources, such as computational power and storage, more efficiently. The platform aims to harmonize user needs with system capacity, ensuring optimal performance, thereby contributing to cost-effectiveness and environmental sustainability.
- Customised **User Experience:** Knowing the unique needs of different user roles, the platform is designed to deliver a tailored user experience. By offering the most relevant features and data in a user-friendly interface, we aim to boost productivity, user satisfaction, and ultimately, patient outcomes. For example a clinician-researcher might only have access to resources for which there is a minimal quality threshold met, whereas academic researchers might see the entire range of models regardless of their quality/credibility.
- **Data Privacy and Security:** In the era of data breaches, the need for robust data privacy and security mechanisms cannot be overstated. Well-defined user profiles are the foundation for implementing stringent access control, ensuring that each user group only accesses data and functionality relevant to their roles, thereby safeguarding sensitive information.

- **Compliance with Regulations:** Healthcare data is governed by strict privacy and security regulations. Well-defined User Profiles and User Roles enable the platform to enforce compliance effectively, ensuring user access aligns with legal requirements, such as GDPR in Europe.
- **Streamlined Collaboration:** By facilitating clear distinctions and interactions among user roles, the platform streamlines collaboration among healthcare professionals, researchers, and regulatory agencies. This is a step towards a future of improved patient outcomes and accelerated medical advancements, driven by efficient information exchange and mutual growth.
- **Monitoring and Accountability:** Clear user profiles form the basis of effective monitoring, aiding us in tracking user activity and identifying potential areas for improvement. This accountability is essential for system integrity, encouraging responsible usage, and enabling swift responses to unauthorised or suspicious activities.
- **Scalability and Flexibility:** As the concept of VHTs expands, well-defined user profiles will ensure the platform's ability to scale efficiently and adapt to new user groups and requirements. This flexibility is key to the platform's long-term viability, as it allows us to evolve alongside the ever-evolving healthcare landscape.

Each of these elements forms an integral part of our strategy, and they collectively contribute to our overarching goals.

Recognizing the diversity and complexity inherent in the stakeholders backgrounds, experience and intended use, we understand the importance in well-defined ***user profiles*** **and** ***roles*** will play in system efficiency, data security, regulatory compliance, and user experience.
In the VHT infrastructure a **user profile** is defined as a collection of settings and information associated with a user. It contains critical information that is used to identify an individual, such as their name, age and individual characteristics such as affiliations, background knowledge or expertise. The profile should be linked to an external IDP (Identity provider) and thus the authentication should be performed with trusted third party services (such as trusted National Identity Provider services in current EU states or an European wide citizens identification system such as the personal digital wallet for EU citizens and residents thanks to the trust framework created by the eIDAS Regulation in the long term.) *User profile does NOT distinguish the role of the user* in the system. User profiles are adaptable as the individual changes jobs, relocating to other countries, acquiring other knowledge or expertise.
User role is a well-defined collection of capabilities within the system allowing the user to achieve the role's intended purposes. In other words, user roles are clusters of system privileges that are designed to achieve specific goals using the VHT infrastructure. User roles are dynamic as their assignment to user profiles. A user profile can consist of various user roles as a person can have different aims in using the VHT platform in their capacity. New roles can be defined as new paradigms of using the VHT platform emerge. In the initial phase we envision the implementation of three categories of roles in the VHT infrastructure (but necessarily all roles within each category). They are mentioned here in the likely order of their implementation in the infrastructure set-up phase.

Creator/Model Developer/Researcher Category (can upload new models or algorithms, maintain and train existing models or algorithms)
Data Scientist: Access to anonymized data for developing models, as well as access to model training and evaluation tools.
Simulation Engineer: Access to simulation tools, environments (HPC), and relevant data to design and validate virtual human twin simulations.
Model Developer/Owner: Upload new model version, manage existing models (*e.g.* configuration, model access constraints, *etc.*)
Biomedical (Engineering) Researchers: Access to simulation tools, environments (HPC), and relevant data to use, compare and validate simulations to gain insights into human (patho)physiology.

Healthcare Professional Category (Healthcare professionals, Doctors, Specialists, *etc.*)
General Practitioner: (can create new simulations for their patients) Access to their patients' VHTs and relevant medical data to monitor health, diagnose conditions, and recommend treatments.

Medical Specialist: Access to group specific patients' VHTs and related data within their area of expertise (*e.g.*, cardiologists accessing cardiac data) to provide specialised consultation and treatment recommendations.
Medical Researcher (can run cohort analysis): Access to anonymized data sets and research tools to conduct studies and contribute to advancements in medical knowledge.
Medical Educator: Access to VHTs for teaching and training purposes, allowing students and professionals to learn and practice in a simulated environment.

Patients/Citizens Category
Patient (Access to their own VHT, personal medical data, and treatment options, enabling them to better understand their health and make informed decisions about their care.)
Patient Advocate: Access to a specific patient's VHT and medical data with the patient's consent, allowing them to provide support and guidance in healthcare decision-making.
Citizen Scientist: Limited access to anonymized data sets and research tools to contribute to community-driven medical research initiatives.

For the latter category, data access can benefit from ongoing research in terms of personal health data access, such as in the Solid Community Group[72] or the development of personal Health Data Pods[73] (POD: personal online data) in EU member states. These data pods could contain a wide array of health data types ranging from hospital records, such as MRI scans, ECG results, personal genome data and diagnoses, to pharmacy data, and even data captured from wearable devices. The VHT infrastructure will need to foster an environment where personal health data can be securely accessed and effectively utilized (primary use) while retaining patient consent.

5.3.3 Data reuse
The practice of data reuse is a fundamental element as DTH models are envisioned to be integrated at input-output level towards creation of the VHT. Hence, the methodology to be followed involves not only **integrating varied data categories** - encompassing clinical biochemistry, radiological modalities such as MRI and CT scans, genomics, proteomics, and metabolomics - but also **amalgamating cohort-specific data and data specific to individual patients or citizens** (contingent on their informed consent) as well as **simulation outputs created within the VHT platform**. The establishment of such a robust, multidimensional data catalogue is of key importance, capable of propelling the innovative potential of DTH towards the creation of the VHT.
Strategic collaboration with other European data initiatives such as OPENAIRE serves as an instrumental aspect of this approach. OPENAIRE, a strong proponent of the open science paradigm, contributes a multitude of invaluable data sets. The VHT platform will be designed to be harmoniously aligned with the ethos and ambitions of GAIA-X, which advocates for a protected, federated infrastructure that enhances data availability. Similarly, the European Open Science Cloud (EOSC) reflects our ethos for endorsing data sharing and reuse within the scientific fraternity. Through these collaborations, the VHT infrastructure will augment its data assets further and stimulate a culture of open, cooperative investigation and will be an integral component of Europe's formidable digital trajectory.

Acknowledging the criticality of a harmonised and interoperable health data environment, the VHT resonates with the objectives of the **European Health Data Space (EHDS)**, which aims to elevate the accessibility, quality, and application of health data across EU jurisdictions. However, in recognizing the considerable power this confers, we also acknowledge the accompanying accountability. Our unequivocal commitment to data reuse within the VHT operations strictly adheres to European ethical and societal principles. Rigorous compliance with legal regulations and guidelines is an absolute mandate. VHT processes safeguard individual data privacy rights, consent and staunchly uphold principles of data minimization and purpose limitation.

[72] https://www.w3.org/community/solid/
[73] https://we-are-health.be/en

5.3.4 Data transformation services: harmonizing and transcending boundaries

The VHT is to be a robust, consolidated and interoperable environment that embraces diversity at both individual and population levels. We foresee tremendous variety of *data types* covering spatial and temporal dimensions of personal and population level data as well as various *data formats* within each data type. Hence, we envision data transformation services within the VHT platform to be designed to seamlessly transform, reformat and integrate various data sources (with models), helping the users to generate rich, actionable insights. These services are provided by the VHT platform, empower the users to harness the power of diverse data types and formats, and model resources effectively and efficiently.

We foresee the following categories of data transformation services that would be performed by the VHT:

1. Unit Conversion

Within the rich and varied health data landscape, diverse data sources often present information in different units or measurement systems. Our data transformation services are designed to employ standardised techniques to reconcile these disparities, mapping data from one unit or measurement system to another. This harmonisation process allows data captured under specific protocols to be transformed into universally comprehensible units, facilitating seamless integration and interpretation across various models and applications.

2. Format Conversion

Complementing the unit conversion/transformation process is the ability to handle data format conversion. users can convert data captured using one technology into a format compatible with other technologies, without compromising the spatial-temporal scales. This conversion is essential to ensuring compatibility and interoperability across the platform, allowing the most effective use of diverse data sources.

3. Interlinking Models: The Input-Output Perspective

Beyond data conversion, we strive to facilitate the smooth interaction between different models. This involves effectively handling the input and output data of diverse models to enhance their synergistic functioning. In doing so, we leverage the interplay between various models in VHT workflows to generate holistic, reliable predictions and simulations.

4. Personal to Population Level Transitions

One of our data transformation service's unique capabilities would be its ability to fluidly navigate between individual (clustering = 0) and population-level (clustering = 1) data. By transforming and aggregating data from individual profiles, we can generate valuable insights at the population level which are stored as population level aggregates. Conversely, we can draw on these broader patterns to inform and refine our individual-level predictions and interventions when individual level data is scarce or lacking. This dynamic interaction between micro and macro perspectives drives a comprehensive approach to healthcare that respects individual uniqueness while maintaining a population-oriented outlook.

5. Segmentation

Another important transformation service would be in data segmentation of personal health data. This involves dividing the health-related information of individuals into distinct segments based on specific criteria. In this type of transformation we do not change the data object but its grouping. Some examples are :

- *Medical Conditions*: Segmenting individuals based on the specific medical conditions they have, such as diabetes, hypertension, asthma, or cancer. This segmentation is crucial in parameterizing models with the correct individual's data for understanding disease prevalence, treatment effectiveness, and tailored healthcare interventions in a digital twins context.

- *Age Groups*: Dividing individuals into different age groups, such as pediatric, adult, or elderly populations allow the models to treat these segments of the population separately. This segmentation is important as health needs, risks, and treatment options may vary across different age groups.
- *Gender:* Segmenting individuals based on their gender, as certain health conditions or treatments may be specific to a particular gender. For example, breast cancer models are more relevant in females and/or prostate issues in males.
- *Genetic Factors*: Segmentation based on genetic information, such as specific gene mutations or variations. This segmentation can help model the impact of the individuals at higher risk for certain diseases or guide personalised treatments.
- *Lifestyle Factors:* Segmenting individuals based on lifestyle factors like smoking status, physical activity levels, diet patterns, or alcohol consumption. This segmentation can assist in assessing (including or excluding) health risks, modelling their effects and designing targeted interventions.
- *Electronic Health Record (EHR) Data*: Segmenting individuals based on their medical history, including hospital visits, medication usage, lab test results, or surgical procedures. This segmentation allows models to know which data is available from the healthcare providers and can create insights into patient populations, disease progression, and treatment outcomes.
- *Treatment Response:* Segmenting individuals based on their response to specific treatments or interventions. This segmentation helps in modelling the response, evaluating treatment effectiveness, identifying subgroups that benefit the most, and refining personalised treatment plans.

As detailed in Section 5.4.1, our models are more than just predictive tools; **they also serve as data transformation services**, producing novel data that is stored within the platform with full provenance details. Thus, our data transformation services promote a harmonised, interoperable platform that utilises health data to its fullest potential, irrespective of its origin, format, or intended scale of use.

5.4 Models

5.4.1 Model as data transformation services

We have defined a Digital Twin in Healthcare (DTH) as a computer model that predicts quantities of interest (QoI) necessary to support decision-making within a Context of Use (CoU) in healthcare. These DTHs can be generic, population specific or even fully personalised subject specific. They could *e.g.* model a cerebral aneurysm and the QoI could be the risk of rupture of that aneurism in the next year. The CoU could be fully personalised, or for individuals in a sufficiently stratified population (*e.g.* caucasian females over 45 years of age without other conditions), or much broader in a large population of all individuals that present themselves with a diagnosed cerebral aneurysm.

Let us now introduce a few generic concepts that will help in sharpening the role of models, data, and models as data transformation services. It starts with the notion of a *system* that we study and that we may want to change. The system could be the example of a cerebral aneurysm as a pathology that can present itself in humans. The intended CoU will further define the system, it could be the overall pathologic condition of a cerebral aneurysm in a full population, or the specific aneurism of a specific individual as presented at a specific point in time.

We now define a system S as a potential source of data. To learn about the system, we experiment with it. For instance, with medical imaging we can observe the cerebral aneurysm and extract data from it (location, size, shape). We define an experiment E as the process of extracting data from a system by exerting it through its inputs. With this we can now define a model M for a system S as anything to which the same experiment E can be applied to answer questions about S. Going back to the example of the cerebral aneurysm, an experiment that could be done, but not in the real person, is to see under what pressure loadings the aneurism would rupture and then asking what the probability would be for such loading to appear in the next year. Such an experiment could be performed on the model M, noted $E(M)$, and the expectation is that the resulting data of $E(M)$ is close enough to the resulting data of E applied to S, noted $E(S)$. More precisely, the output data error of $E(M)$ compared with $E(S)$ should be sufficiently small and clinically acceptable in the intended CoU. When this is the case we say that the

model *M* of the system *S* is valid in the intended CoU (where we imply that a clear definition of the experiment *E* is part of the CoU).

A DTH is a mathematical model, which, in the vast majority of cases, is not amenable to analytic solutions. So, performing an experiment on a DTH will require a simulation on a computer. This notion is captured by defining a DTH as a computer model, so a mathematical model that will be solved on a computer to experiment with it. Do note that we need to make an important distinction between the model description (in terms of the mathematical formulation and its coding as a computer program) and the experiment description (so, setting up the specific inputs to the computer program that encodes the model). Also, note the danger of this separation. It is very easy to apply an experiment to a computer model for which it is not valid (*e.g.* by setting input parameters beyond the range for which the model was validated).

Establishing the validity of a model, so if *E(M)* is close enough to *E(S)* within the CoU for the specific QoI, is far from trivial. Remember that a model is always related to the tuple system and experiment. Model validation *always* relates to an experiment to be performed on a system (*E(S)*), which should be very clearly defined by the QoI and the CoU. Demonstrating validity is a key element in the credibility assessment of DTHs (see *e.g.* section 5.2.5). As noted in section 4.2.4 (in relation to barrier 5) we will rely on the ASME VV-40:2018 as a standard for emerging regulatory pathways in Europe in relation to assessing credibility of DTHs.

A final relevant notion is that a model can also be qualified as a system. So, a model of system *M(S)* is by definition a source of data and therefore also a system *S' = M(S)*, and we could create a model *M'(S')*. For instance, a very complex and computationally demanding multiscale DTH could be considered as a system, for which a computationally much cheaper data driven surrogate model could be constructed. Note that we now create hierarchies of models and related experiments, so we must be very careful in keeping track of the original system, *e.g.* a specific person or a stratified population with a certain pathologic condition, for which a model of a model is supposed to produce data that resembles the data for the original system. Clearly, well-defined protocols are needed to keep track of such hierarchies of models.

From the infrastructure point of view, a computer model is then viewed as a data transformation service, and a DTH is then composed of one or more of those services. In Figure 3 the basic idea is presented. Input data would be a set of model parameters, a set of boundary conditions and a set of initial conditions . The model M_{CoU}, valid in a specific CoU, then transforms the input dataset $X=\{\alpha,\beta,\gamma\}$ (defining the experiment *E* to be performed on the model) to the model output *Y*, from which in turn the QoI is extracted, $QoI = f(Y)$ and $Y = M_{CoU}(X)$.

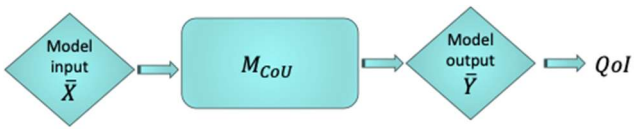

Figure 3: a computer model as data transformation service.

In its simplest form a DTH would be like in Figure 3. In reality, a DTH is usually much more complex, where the model M_{CoU} could be executed many times (*e.g.* for uncertainty quantification), where other data transformation services are needed to prepare the model input (*e.g.* to extract a geometry of a cerebral aneurysm from an MRI image), where different models are chained together, and where additional data transformation is required to get the final QoI from the model output, and any combination of those.

5.4.2 Model as data generation services

As noted above, a model can also be seen as a system itself, and as such becomes a source of data that can subsequently be used for creation of other models. For instance, in the example of the aneurysm, one could aim at creating some reduced order model for the pulsatile flow in the aneurysm, and use the output from the original model that computes these flow fields to obtain *e.g.* a proper orthogonal decomposition on which the reduced order model could be constructed. In this scenario the model is used as a data generation service, and the data is then reused to construct or train other models, possibly relying on recent developments such as *e.g.* Physics Informed Neural Networks. The most obvious reason to do so is when the original model requires huge computational resources, maybe millions of core hours on dedicated tier0 or tier1 HPC, and a computationally cheaper surrogate model is needed when it should be applied in scenarios when fast response is needed (*e.g.* in decision support scenarios), when many instantiations of the model are needed (*e.g.* for uncertainty quantification), or in *e.g.* computer aided design scenarios. Many high fidelity time dependent three dimensional multi-{scale, component, organ} models are usually very compute intensive, and to put them to use for *e.g.* clinical decision support, treatment optimization, or computer aided design of implants, requires creating (machine learned) surrogate models, and the VHT infrastructure should provision ways to do that.

As noted above it is very important to keep track of the CoUs/QoI of the original high-fidelity model. The vision is that output of each run of such high-fidelity simulation is stored in the VHT infrastructure, ready to be reused, possibly by independent third parties and for their own purposes. In such a scenario the third party may not even be aware of the origin of the simulated dataset (the model that created the data) and could use that data as a ground truth for training surrogates. When the CoUs of the simulated data is not kept clearly connected, as metadata, there is the danger of using simulated data beyond its CoU.

5.4.3 Models as data flow orchestrations

We have two basic objects, a set of models M and a set of data object D. They are connected as a directed bipartite graph, meaning that elements of M (so, individual models) take as input elements of D and produce as output elements of D that in turn can serve again as input to other elements of M. Consider e.g. a simple workflow, with 2 models $M = \{m_1, m_2\}$ and five datasets $D = \{d_1, \dots, d_5\}$ chained together as shown in
. One can also view the resulting graph as a coloured graph, with data in orange and models in blue, and each edge in the graph going from colour to another.

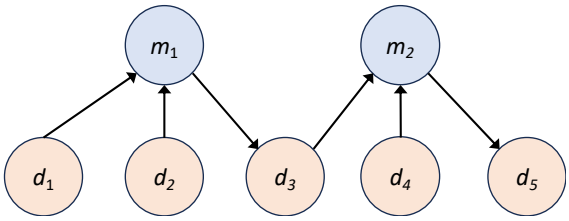

Figure 4: example of workflow graph.

If we now assume that the M and D form the full set of model – and data objects in the VHT, and the set of edges E describing all input-output relations between data and models, we can denote the graph $G_{VHT} = (M, D, E)$ as the formal representation of the collected information available in the VHT. G_{VHT} is a dynamically growing object, as we keep adding data and models to the VHT and link them up to the G_{VHT}.

We should not confuse the (ever growing) G_{VHT} with the final model of the full pathophysiology of a human being. G_{VHT} may contain many different models that compute more or less the same QoI but in different CoUs,

or at different spatio-temporal scales, or at different clustering, or operating at different levels of credibility. However, G_{VHT} does contain many different DTHs, expressed as workflows. In the example of

, we could imagine that this specific small graph is embedded in the G_{VHT}, and view this as a workflow with three input data objects $\{d_1, d_2, d_4\}$, one intermediate dataset $\{d_3\}$, one output data set $\{d_5\}$ and two models $\{m_1, m_2\}$. We could write $G_{DTH} = (M_{DTH}, D_{DTH}, E_{DTH})$ and $G_{DTH} \subset G_{VHT}$.

So far G_{VHT} is a growing but otherwise static object. However, models execute, produce new data, if new data is available, other models could execute, as described in section 5.2.4. The idea would be that the models in the VHT as it were scavenge on data, and once new data becomes available for them, they execute (so, a dataflow execution model). To add such dynamic model execution to the whole picture we can resort to Petri Nets, which is "a directed bipartite graph that has two types of elements: places and transitions. A place can contain any number of tokens. A transition is enabled if all places connected to it as inputs contain at least one token", and "a transition of a Petri net may fire if it is enabled, i.e. there are sufficient tokens in all of its input places; when the transition fires, it consumes the required input tokens, and creates tokens in its output places."[74] In our case places are data objects, tokens new data items arriving or availability of computing and storage resources, and transitions are the models.

5.4.4 Model classification by context of use

In a grass roots projects, ran through the ISW_CoP, a non-exhaustive list of possible contexts of use for models was established[75]. It use the 3Rs principle (reduce, refine, replace) to structure the exercise, focusing on 3 categories of experiments that benefit from the use of computer modelling and simulation: (i) preclinical/in vitro/ex vivo experiments, (ii) preclinical animal experiments, (iii) clinical human experiments. For 31 of the identified contexts of use, preliminary evidence has been provided in the literature. Below, some examples are provided of CoUs that are highly relevant for the VHT.

- **Diagnosis**: Indicate potential biomarker for disease: (*e.g.* digital biomarkers for cerebral vasospasm [76]) or substitute invasive measurement of disease biomarker (*e.g.* pressure gradient across coronary stenosis[77] or right heart catheterisation for diagnosis of pulmonary hypertension[78]).
- **Prognosis**: predict risk of future adverse event (*e.g.* hip fracture[79] or aneurysm rupture[80])
- **Stratification**: determine, justify and/or confirm eligibility or exclusion criteria for proper patient or treatment selection
- **Treatment**: use as clinical decision support systems. Recommend pharmacological treatment course for a specific condition (*e.g.* management of glycemic levels in ICU patients[81]).
- **Development of medicinal products**: inform clinical development decisions (e.g.)
- **In silico clinical trials**: Use digital biomarker as primary endpoint in the clinical trial (e.g.); Use adaptive clinical trial designs, using the results of in silico trials on virtual patients as priors

5.5 Integration of resources

5.5.1 Identification of possibilities for integration

The main resources, data and models, can be integrated in a number of ways, and can then be mapped to the computing, storage, and networking resources. Data can be grouped together, pooled, to form richer datasets, maybe forming a new DOT. Models can be integrated into multiscale / multiorgan

[74] https://en.wikipedia.org/wiki/Petri_net
[75] https://pubmed.ncbi.nlm.nih.gov/34161248
[76] https://doi.org/10.1016/j.jbiomech.2019.04.019
[77] https://doi.org/10.1093/ehjdh/ztac045
[78] https://doi.org/10.1086/686020
[79] https://www.ansys.com/advantage-magazine/volume-xv-issue-3-2021/simulation-and-high-performance-computing-reduce-fracture-risk-in-osteoporotic-patients
[80] https://doi.org/10.1007/s10237-020-01351-2
[81] https://doi.org/10.1016/j.cmpb.2010.12.008

models. And integrating chains of models and data together where output data of one models serves as input data to another model results in workflows, that are considered the third basic resource in the VHT.
Likewise computational and storage resources can (dynamically) be pooled together to provide the required infrastructure to interrogate data, store (new) data, and to execute workflows. This will be further discussed in section 6.6.

5.5.2 Integration of multiscale models

Multiscale models have model features at multiple scales of space and/or time. The Multiscale Modelling and Simulation Framework (MMSF) was defined[82,83]. It is based on the Multiscale Modelling and Simulation Language (MMSL)[84], which describes the coupling of many sub-models to the multiscale model.
One proposal is to integrate machine learning and multiscale modelling[85], where machine learning explores big parameter spaces for the identification of correlations and the multiscale modelling predicts the system dynamics and identifies causality. In such combined models system validation is of prime importance to avoid that the ML algorithm leads to non-physical resp. non-biological solutions.
Prospectively, markup-languages permitting a unique description of models should describe all mechanistic models, including even agent-based models of individual cells that are becoming increasingly sophisticated. MultiCellDS or MultiCellML may be promising steps into this direction.
Tightly coupled multiscale models, where models interact in cycles (e.g. a microscale model providing input data to a macroscale model, that in turn at a next time step provides input data the microscale model, etc.) require dedicated coupling frameworks, as discussed above, and from the point of view of the VHT are considered a new model. In contract, loosely coupled multiscale models do not have such cyclic feedback loops and can be composed as a workflow, and are there viewed as a workflow resource.

5.5.3 Workflows

Workflows allow for the fast deployment of the VHT in different computing architectures (such as Cloud and HPC, see section 5.6.3) and enables the reproducibility of use cases across different users.

Technically, the use of workflows greatly **facilitates the deployment** of complex and large use cases **in different computing facilities** and the **generation of intermediate data structures** in case the simulation has execution errors or perturbations (as energy outages). They also enable **repeatability and reproducibility** of use cases while providing a way to integrate provenance of data and privacy by allowing for federated workflows across multiple sites. In the VHT, automatically keeping track of provenance, certainly in relation to the credibility of the output data of a workflow, is a key requirement.
Workflows **facilitate user understanding** of the overall DTH without the need of delving into detailed model documentation and metadata and lowering the learning curve to be able to simulate use cases by themselves.
For example, a workflow recipe for a complex use case for modelling brain tumours combining different types of models, data sources, and output data for a commonly used workflow manager can allow researchers to seamlessly deploy exactly the same use case in the Cloud (*e.g.*, AWS and Google Cloud) or in different high-end HPC clusters (*e.g.*, LUMI, MareNostrum) with minimal amount of work.

Even more so, the use of workflows usually but not necessarily involves **the division of the work in containerised pieces** (such as Docker, Apptainer (former Singularity). This compartmentalization allows for the decoupling of the complexity of each part of the workflow and its usability in complex use cases. Likewise, these containers facilitate benchmarks of different tools aimed at the same or similar tasks by solving unit tests to study *e.g.* the validation, verification and uncertainty quantification (VVUQ) of each model and implementation and also allow for the definition of federated workflows

[82] Borgdorff *et al.*, 2013
[83] Chopard *et al.*, 2014
[84] Veen, Hoekstra, 2020
[85] Alber *et al.*, 2019

by clearly dividing the work of each software piece and assigning them a given computing cluster (one part in the cloud, another on-site *etc.*). Also automating this process, or creating a demand/supply market to for semi-automated execution of complex DTHs, in relation to the dataflow mechanisms introduced earlier in section 5.4.3, can be facilitated by advanced VHT workflow engines.

Despite this, the organisation of use cases in containers and workflows is a **clear overhead on the researchers' side** and currently there are very few ways that this conversion can be leveraged. There are a myriad of different **workflow managers** in the biomedical domain with slightly different focuses; some of them open source, while others commercials (Keppler, Nextflow, Galaxy, cwltool, Apache Taverna, Cromwell, Snakemake, KNIME, PyCOMPSs, Rbbt.
For these reasons, the community has been working on delivering **standardised workflow descriptions** such as the Common Workflow Language (CWL) and the Workflow Description Language (OpenWDL) of Cromwell to facilitate the deployment of workflows' recipes across managers. Likewise, workflows registries have been established to facilitate the discovery and re-use of workflows in an interoperable way (Dockstore, WorkflowHub), as well as frameworks providing tools for construction and execution of workflows (WfCommons) and backend programs for the reproducible execution of standard workflows (WfEx-S or Workflow Execution Service).

5.5.4 Bidirectional communication with users (including knowledge generation)

To be completed

5.6 Infrastructure

The **infrastructural requirements** of a VTH platform are a **direct consequence of the requirements** of the underlying computational models and data storage facilities. In order to provide the services envisioned in this document, both the repository and platform need to be able to access a variety of computational resources, and operate across organizational boundaries, including in an international context. Enabling such operation calls for resolution of a number of issues of a legal, organizational and technical nature.
A useful primer to European infrastructural standardization efforts is provided by the Rolling Plan for ICT Standardisation. This is a live document, curated by the European Commission in collaboration with the European Multi-Stakeholder Platform (MSP) on ICT Standardisation. It lists all the topics identified as EU policy priorities where standardisation, standards, or ICT technical specifications may play a key role in the implementation of the policy.
The remainder of this section addresses specific ICT considerations applicable to each subcomponent of the VTH infrastructure – specifically, the data repository, simulation platform and computational resources.

Regarding the **Repository and the Platform**, the selected software solution should act as a one-stop shop for the scientific community to discover, share, design and use Virtual Human Twins (VHTs).
The suggested solution will consist of three interconnected components (Figure 5): a Catalogue for storing metadata for any Research Objects (ROs) related to VHTs (models, datasets and simulation definitions *etc.*), a Repository for storing data related to ROs registered to the Catalogue, a Platform to perform data analysis and simulation using ROs from the catalogue and the repository. The Platform will be composed of several tools including workflows execution engines, simulation and machine learning frameworks.

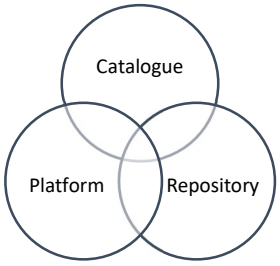

Figure 5: schematic overview of the main components of the VHT infrastructure.

5.6.1 Repository

As detailed in Section 5.2, a data repository is envisioned as a space where VHT data assets can be uploaded, stored, discovered and shared by users. In this context "data" refers not only to data objects (*e.g.* input files or results of VHT simulations), but also the models themselves, stored in the form of software artifacts. The European community has had some success in building such cross-organizational repositories – as evidenced by the EOSC Marketplace[86], the EUDAT project or the Fenix infrastructure[87] – and we believe that a VHT repository framework should build upon the achievements of such initiatives, while also extending their capabilities to provide support for a VHT-specific metadata storage model. In this sense, the repository should provide support for the following:

- **Catalogue services**, enabling discoverability of resources,
- **Secure role-based access** (see section 7.2 for a discussion of user roles)
- **Dedicated storage resources** for unstructured binary data, structured databases and model repositories (with CASE-specific features, such as are provided by modern code versioning tools).

We also regard it as crucial for any actionable VHT data repository to provide a consistent set of programmer's interfaces (APIs) through which content could be accessed programmatically rather than directly by human actors. Such a feature would facilitate development of automated computational workflows and enable large-scale VHT simulations.

5.6.2 Simulation platform

A simulation platform is regarded as complementary to the data repository discussed above. This layer comprises **computational services deployed and accessible** throughout the European ICT ecosystem. While referred to as a singular "platform" it can, in fact, comprise **multiple autonomous services** which perform processing of data stored in the repository. The platform – or its constituent services – should be able to interact with the previously described storage resources, and should enable deployment of computational models across organizational boundaries. To this end, we believe that VHT models should follow a uniform packaging scheme, such as is provided *e.g.* by containerization frameworks (Docker, Apptainer, *etc.*)

We also believe that the added value of a VHT platform rests in the "three Rs": **Repeatability, Replicability, Reproducibility**. Particularly in the context of medical simulations, the need to ensure validation and repeatability of computations, as well as to track their provenance, is particularly relevant, and any computational tools developed on top of the VHT data repository should enable such information to be tracked and accessed where necessary using widely-accepted packaging schemes.

The simulation platform should also expose dedicated interfaces to various groups of stakeholders, which correspond to the various user roles listed in section 7.2 of this document. As remarked above, these interfaces (such as workflow management panels, data input forms, visualization tools, administrative mechanisms *etc.*) can be implemented by autonomous tools, as long as standardized means of communication are maintained – by agreeing upon a set of shared APIs.

VHT platform should **analyse, simulate, visualize and process Research Objects**, while, at the same time, should also **facilitate the management of the results and the interaction** between the users. As already mentioned, it will offer several software services (web apps, APIs, Jupyter notebooks, workflow engines, VDIs) and will support several types of computational resources discussed next.

5.6.3 Computational resources

Computational resources form the bottommost level of the VHT technology stack. These can take on the form of "bare metal" computational infrastructure, but in the modern age they are more frequently encountered in virtualized form, which facilitates large-scale management and supervision of complex ICT infrastructures. In general, we perceive support for the following broad types of computational resources:

- **"Classical" HPC** – embodied by computing clusters. These are useful, and often outright required, when performing complex computations or running large-scale parameter sweeps or cohort studies. HPC infrastructures are usually operated by dedicated computing centres, many of which work on a non-profit

[86] https://marketplace.eosc-portal.eu
[87] https://fenix-ri.eu/infrastructure/resources/available-resources

basis and receive public funding in order to support (among others) scientific research projects. Thus, they are often well integrated with the broader European data and computing infrastructures.

- **Cloud services** – a somewhat more modern take on access to computing services, cloud infrastructures are typically made available as sets of virtual machines, procured on an as-needed basis and particularly useful in supporting the so-called "long-tail science" (which includes many types of VHT simulations and data processing tasks) where "classical" HPC resources are not regarded as necessary. For the most part, cloud services are the domain of commercial providers – such as Amazon, Microsoft or Google – although dedicated scientific infrastructures do exist. Cloud services require different access mechanisms than HPC clusters, and both are well suited to deployment of containerized software – which is why we believe that containerization is an important aspect of deploying software services in the context of a heterogeneous VHT infrastructure.
- **End-user devices** – while not normally regarded as part of computing infrastructures, they play an important role in any self-contained infrastructure, enabling delivery of useful information to end users. In the medical context, given the broad range of potential VHT simulations and use cases, issues related to development of suitable access interfaces for end users (many of whom are not IT experts) should be taken into account.
- **Edge computing** - also referred to as fog computing, the concept of a processing node which is topologically proximate to end-user devices is useful in situations which call for low latency and process real-time data, such as those generated by sensors. In such circumstances, for applications deployed in a medical care setting, it might be useful to deploy a local compute node, capable of processing urgent requests, while more computationally heavy tasks are delegated to external HPC resources. Such local nodes are often operated - on a limited scale - by medical care providers, and can be leveraged in the context of VHT applications for persons receiving treatment at a given location.

5.6.4 Recommendations for VHT Infrastructure

HPC and Cloud infrastructures offer computational and storage resources of the highest class to users and projects, based on scientific or strategical merit, and essentially free of charge. Getting access to, deploying and managing software on, and running simulations on these resources each come with their own set of challenges, as has been shown above.

To build a sustainable and manageable infrastructure on top of these (heterogeneous and changing / evolving) resources will require

- A simple(r) usage model, based on
  - Long running (strategic) grants as opposed to the typical project/grant application cycle, wherever possible
  - Service access instead of individual user access, with the resource provider delegating part of the user and access management to the service. This would make scientific HPC and cloud offerings much more accessible for a wider user base, and in many cases make using the service possible to
- Software deployment and execution heavily based on containers
  - Need to define the type of resources an application / workflow requires
- Access APIs should be in place wherever possible (*e.g.* UNICORE), that allow flexible federated authentication and some level of abstraction

5.7 Conclusion

TBD

6 Standards, regulations, legal, ethical and social aspects of the virtual human twin

6.1 Introduction

TBD

6.2 Regulatory science and Standards

The current draft of the Good Simulation Practice consensus document[88] provides in the appendix a systematic review of all regulatory documents relevant to the in silico methodologies used to evaluate new medical products (In Silico Trials) trials.
The regulatory science for in silico medicine is largely in the making. Until recently, the European regulatory framework for medical devices did not explicitly recognise that **software could be a medical device or be part of a medical device**. The idea was introduced with the 2007 revision of Directive 93/42. Still, it is only with the new Medical Device Regulation (hereinafter MDR, Regulation (EU) 2017/745) that it is recognised as the Medical Device Software category (which in other regulatory systems is called **Software As Medical Device**, SaMD).
Predictive software is first regulated in the context of In Silico Trials, where the predictive model is used to refine, reduce or replace experimental studies (whether done in vitro, on animals, or on humans). In 2016 the FDA Centre for Devices and Radiological Health (CDRH) published a guideline entitled "Assessing the Credibility of Computational Modelling and Simulation in Medical Device Submissions", followed in 2018 by the publication of the **ASME VV-40:2018** technical standard, entitled "Assessing Credibility Of Computational Modelling Through Verification And Validation: Application To Medical Devices". This opened the possibility of producing evidence of safety or efficacy for new medical devices in their regulatory process. Soon the credibility assessment approach proposed by the VV-40 was also recognised by the FDA CDRH as valid for the certification of SaMD with predictive capabilities. In the European Union, notified bodies have already accepted in silico evidence to reduce in vitro experimentation. However, the start of standardisation activities by ISO/IEC aimed to develop standards with a scope similar to the VV-40 opens up a future where also for the CE marking of medical devices based on evidence provided by in silico methodologies, or the CE marking of predictive SaMD, will be possible using a **harmonised standard**. And meanwhile, it is reasonable to expect the EU notified bodies to accept transitorily evidence of credibility based on the ASME VV-40:2018.
In published papers, staff members of the FDA Center for Drug Evaluation and Research (CDER) and of the European Medicine Agency (EMA) suggested that the VV-40 could be used to demonstrate the technical validity of a new in silico methodology to be used as a drug development tool. However, how to demonstrate the clinical validity of in silico drug development tools is still debated.

What follows is an overview of the currently available standards and regulatory guidances. The final choice of the optimal standards to be used will depend on the choices made in the tech stack during the implementation phase of the VHT.

6.2.1 Standards for data formats, data integration and data input into models

The ISO 20691:2022[89] standard "Requirements for data formatting and description in the life sciences" requires that the data in the life sciences shall be FAIR (Findable, Accessible, Interoperable and Reusable). Therefore, the data used in models and simulations of the human digital twin should at least encompass the information described in the minimum information standards and follow the FAIR[90] principles. For being FAIR, the data must have a unique ID, must be linkable, must have an assigned

[88] https://insilico.world/sito/wp-content/uploads/2023/05/Position-Paper-GSP-R6.pdf
[89] https://www.iso.org/standard/68848.html#:~:text=This%20document%20specifies%20requirements%20for,human%20biological%20research%20and%20development.
[90] https://www.go-fair.org/fair-principles/

licence and must be annotated by metadata describing for instance the disease, the tissue, the cell type and the used modelling parameters.
Other requirements described in ISO 20691:2022 are about the consistent formatting and documentation of data, models and metadata as well as the requirements concerning storing, sharing, accessing, interoperability and reuse of data, models and metadata in the life sciences.
The ISO 20691:2022 standard acts as a reference framework or hub standard for other life science data and integration standards. It describes the requirements and rules for applying standards for formatting, description and documentation of data types in the life sciences and contains a catalogue of criteria and requirements for interoperable life science data formats and semantic data description standards.
The annex A of ISO 20691:2022 contains a list of common recommended formats for life science data and the annex B contains a list of minimal reporting standards for data, models, and metadata. This set of metadata describes the context of the data and models.
The ISO 20691 FAIRSharing collection[91] contains a listing of these standard formats, which should be used to represent life science data and to annotate them by ontology terms. This list encompasses data formats for -omics, biochemical and molecular biology methods as well as formats for biological imaging and for computer models of biological systems. That FAIRSharing collection is actively managed and maintained, so that it is always on the last stand.
A further standard is ISO/TS 9491[92] "Biotechnology — Recommendations and requirements for predictive computational models in personalised medicine research", which consists of two parts:
Part 1: Guidelines for constructing, verifying, and validating models.
Part 2: Guidelines for implementing computational models in clinical integrated decision support systems.
Another ISO standard is ISO 4454:2022[93] "Genomics informatics – Phenopackets: A format for phenotypic data exchange", which standardised the description of phenotypic information.
The ISO 23494 series[94] "Biotechnology – Provenance information model for biological material and data" describes how to document the provenance information for biological data and samples.

Data integration

Data integration combines data from different sources like Electronic Health Records (EHRs), life-style data, environmental and registry data, molecular -omics data, laboratory data, imaging data, biosignal, intensive care vital sign data and other medical relevant sorts of data. Data integration can be done either on the individual level or on the variable level.
Individual level integration is required for personalised models and means that all available data for an individual patient are brought together for the model building and simulation tasks.

It was demonstrated that it's possible to integrate several modelling frameworks (agent-based models, ordinary differential equations, stochastic reaction systems, constraint-based models, solid-body physics and spatial diffusion) into a composite model. But a standard for such an integration of different models into a multiscale model is still missing, even though there are some proposals for such integrative frameworks [Agmon et al., 2022 and Masison et al., 2021].

The ISO/14199[95] standard "Health informatics — Information models — Biomedical Research Integrated Domain Group (BRIDG) Model" is a domain model for data interchange that enables semantic interoperability and intends to bridge between biomedical, clinical research and routine healthcare data. It can be seen as a meta-standard and can be used as a starting point and blueprint for the integration of modelling and simulation data with other health data and how to make them interoperable.

[91] https://fairsharing.org/3533
[92] https://www.iso.org/standard/83516.html
[93] https://www.iso.org/standard/79991.html
[94] https://www.iso.org/standard/80715.html
[95] https://www.iso.org/standard/83433.html

Harmonised access to data across resources

Currently there are no harmonised strategies for data access, since the data protection laws are varying between the countries and can even vary from federal state to federal state as in Germany. What shall be used are strategies to control the access to restricted (*e.g.*, person-related) data by researchers. Such harmonised Data Access Agreements (hDAAs) are still missing now.

Standards for medical imaging

For medical imaging there are three important standards. The first one is the Digital Imaging and Communications in Medicine (DICOM), which is the most used data structure in health care, and which stores the data as 2D layers. The Brain Imaging Data Structure (BIDS) is a group of standards defined by the International Neuroinformatics Coordinating Facility (INCF) and is mainly used in neuroscience/neuroimaging research. For several special imaging techniques extensions to BIDS are available, which contain mainly metadata describing features for these special techniques. A third group of standards were defined by the Neuroimaging Informatics Technology Initiative (NIfTI) and stores the data in a true 3D volume format. Table 3 in the annex gives an overview about these formats and their extensions / descendants.

Standard formats for electro- and neurophysiology, biosignal and vital sign data

Since the annexes of ISO 20691 and ISO/TS 9491 list mainly systems biology formats, in the following some formats for physiology / biomedicine are mentioned. There is a plenitude of formats in the area of electrophysiology, biosignal and time series of vital sign data (*e.g.* body temperature, systolic and diastolic blood pressure, pulse rate, respiration rate, oxygen saturation, BMI…) recorded in intensive care. Many of them are proprietary formats defined by medical device manufacturers. Therefore, in Table 4 of the annex only formats supported by Standard Defining Organizations (SDOs), community standards, or formats, which are at least de facto standards, are listed.

Other models

Models, which are not stored in a standard format and encoded in a scripting language, *e.g.*, in Matlab / Octave, Python, R, Julia scripts, … can be executed on the platform either if they have an interface that allows the import/export, respective conversion into one of the aforementioned standard formats is possible, or if the scripts are directly executed inside the workflow. Otherwise, if the scripts are "injected" into the workflow, they can be generically packed into a Binary FHIR resource.

For some kinds of models, *e.g.*, cellular automata there are no explicit standards available, but at least some of these models can be described by using the MultiCellML and Morpheus standards. For others, like Petri net models there is the ISO/IEC 15909 standard, but this standard is not biology specific.

Minimal clinical core datasets

To optimise the research and patient treatment, some minimal clinical datasets, which are specific for a disease were defined. For example, for cancer there is the mCode[96] (minimal Common Oncology Data Elements) HL7-STU (standard in trial use) available. It describes the minimum information, which should be collected in the electronic health records of cancer patients. Also, for some other diseases, *e.g.*, for multiple sclerosis[97], Parkinson disease and stroke, the definition of minimal clinical datasets is available (see common data elements catalogue[98]).

For some diseases XML-based standard file formats based on such disease-specific common data models for capturing the relevant information are available. Examples are cMDX (Clinical Map Document based on XML) for prostate cancer or BRCAPRO for breast cancer.

Representation of AI-/ML-based models

For AI-based models there are no standardised file formats for the input data available. Since these models are based on big data, they often use hierarchical (.hdf, NetCDF), row-based (.csv, .json),

[96] http://hl7.org/fhir/us/mcode/
[97] https://medical-data-models.org/19067?form-lang=en
[98] https://www.commondataelements.ninds.nih.gov/cde-catalog

columnar (*e.g.* Apache Parquet) or other high-performance (*e.g.* Apache Iceberg) formats as input format for the data used for AI-/ML-based learning.
Most machine learning frameworks define their own model exchange format. For a standardised and interoperable representation and deployment of machine learning and deep learning models one of the Open Neural Network Exchange (ONNX), Portable Format for Analytics (PFA) or Predictive Model Markup Language (PMML) standards can be used. Since PFA is more flexible than PMML and easier to extend to new algorithms and model types and also integrates necessary pre- and post-processing steps, PFA or ONNX shall be used.
The structure and parameters of trained neural networks can be stored and exchanged with the Neural Network Exchange Format (NNEF) defined by the Khronos group.
IEC SC42 defines standards for artificial intelligence and the use of synthetic data for artificial intelligence methods[99].

Genomic sequence variants

For encoding of genomic sequence variants, one should prefer a data format mentioned in the ISO 20691 document, if applicable. An overview about sequence variation formats is given in Table 5 in the annex.
In molecular tumour boards and computational oncology such formats for sequence alignment, genome and copy number variation and DNA methylation patterns are used. These are often stored in the NCI Genomic Data Commons (GDC) or the Portable Format for Bioinformatics (PFB) format. PFB is based on the Apache Avro serialisation format and able to encapsulate the data model, the data dictionary, the data itself and controlled vocabularies in one file [Lukowski et al., 2023].
BioCompute objects (BCO) is the IEEE 2791-2020 standard for documenting workflow steps of bioinformatics analyses of next-generation (NGS) and high-throughput screening (HTS) and is used for regulatory submissions to the FDA.

Finite element modelling (FEM)

For mesh-based finite element modelling (FEM), boundary element method (BEM) or finite volume modelling (FVM) often the .vtk (Visualization ToolKit) output file format is used. For the FEM input files FieldML may be used. For the generation of 3D-models from DICOM data, *e.g.*, for computer guided implant surgery, standard tessellation language (STL) files shall be used, which allow the printing of 3D models.

International Patient Summary (IPS)

International Patient Summary (IPS) is a set of basic patient-related physiological and clinical data and the ISO 27269:2021, EN 27269:2022 and CEN/TS 17288:2020 standard. It comprises data about medications, allergies / intolerances, problems, immunizations, results and procedures for a specified patient. It is a joint standard of five standard defining organisations (CEN, HL7, IHE, ISO and SNOMED) and actively supported by the Global Digital Health Partnership (GDPH) and the World Health Organisation (WHO). It can be implemented either by HL7 V3 Clinical Document Architecture (CDA), with HL7 Fast Healthcare Interoperability Resources (FHIR) or according to the Integrating the Healthcare Enterprise (IHE) IPS profiles.

Phenotypic data exchange

Phenopackets[100] version 2 is a Global Alliance for Genomics and Health (GA4GH) and ISO 4454[101] open standard for sharing disease and phenotype information of a patient/sample, especially in the context of diseases, especially for rare diseases, cancer, but also for common diseases. A Phenopacket links detailed phenotypic descriptions with disease, patient, and genetic information, diagnosis, and treatments.
The phenotypic features are *e.g.*, signs, symptoms, laboratory and imaging findings, behavioural manifestations, or the results of physiological tests.

[99] https://iec.ch/blog/isoiec-report-address-synthetic-data-techniques-used-chatgpt-and-other-ai-tools
[100] http://phenopackets.org
[101] https://www.iso.org/standard/79991.html

The Phenopackets schema is specified as a protobuf schema[102]. These protocol buffers are platform- and language-neutral, so that serialisation code for several programming languages can be generated easily. Therefore, Phenopackets is a standard for the language- and platform-neutral exchange of phenotypic information.

Other data format types

If one has specific modelling requirements, then other data format types than the ones listed above are needed. Examples are FASTQ, a sequencing format with quality information. SAM (Sequence Alignment Map), BAM (Binary Alignment Map) and CRAM (Compresses Reference-oriented Alignment Map) are file formats for alignments.

For microscopy data OMEX (Open Microscopy Environment XML) is a standard for storing acquisition parameters, annotations, and image analysis results of microscopy experiments and BDML (Biological Dynamics Markup Language) is a standard format for storing spatiotemporal dynamics of molecules and cells in live imaging data.

For data from wearable health devices and trackers, *e.g.*, weighing scales, blood pressure monitors, biosensors like blood glucose monitors, smart health watches, mobile ECG devices for arrhythmia recognition… the ISO / IEEE 11073 standard for personal health devices (PHD) shall be followed.

6.2.2 Standardisation of modelling

The modelling process itself shall follow the ISO/TS 9491-1[103] technical standard "Biotechnology - Recommendations and requirements for predictive computational models in personalised medicine research - Part 1: Guidelines for constructing, verifying and validating models".

In order to standardise the modelling process, all data, metadata, models and simulation results shall be documented according to the FAIR[104] and ALCOA[105] (Attributable, Legible, Contemporaneous, Original, Accurate) principles. Therefore, one shall use standard formats with semantic annotations based on defined ontologies.

ISO/TS 9491-1 specifies requirements and recommendations for models used for research purposes in the field of personalised medicine, *i.e.* computational models used in routine clinical, diagnostic or therapeutic purposes are excluded.

In detail ISO/TS 9491-1 contains specifications for:

- the design, development, and establishment of predictive computational models.
- the set-up, formatting, validation, simulation, storing and sharing of computational models.
- data used to construct or be required for validating such models.
- formatting, descriptions, annotations, interoperability, integration, access and provenance of such data.

To standardise the modelling process, the data going into a model shall be integrated and the modelling results shall be validated. The integration of data means the systematic combination of different data necessary for the modelling task and belonging to a patient.

The validation means the comparison between the output of the calibrated model and the measured data.

The annex A of ISO/TS 9491 contains information on common standards relevant for personalised medicine and in silico approaches.

Part 2 of ISO/TS 9491 contains recommendations for implementing computational models into clinical decision support systems (CDSS).

Data preparation

The modelling process and the different types of models are described in detail in the ISO/TS 9491-1[106] document "Guidelines for constructing, verifying and validating models", that describes in detail the steps of data preparation for integrating the data into the models:

- sampling the data.

[102] https://developers.google.com/protocol-buffers
[103] https://www.iso.org/standard/83516.html
[104] https://www.go-fair.org/fair-principles/
[105] https://www.eurotherm.com/life-sciences-cpg/data-integrity-life-sciences/alcoa/
[106] https://www.iso.org/standard/83516.html

- data formatting and harmonisation, *e.g.*, lab values have a unit associated with them. This unit can either be mass/volume or mol/volume. Therefore, the values shall be converted to a unique scale. For that the molecular weight of the analyte must be known.
- data description by descriptive metadata, describing for example the context of the datasets.
- semantic annotation of the data, *e.g.*, by annotating genes and proteins with ontology terms.
- definition of a data interoperability framework3.
- data integration, either on the personal or on the variable level.
- adding data provenance information.
- defining who can access the data.

Model types

In general, one can distinguish knowledge-driven top-down models based on prior knowledge and hypotheses on the causal relationships (mechanistic models) and data-driven bottom-up models *e.g.*, Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) models.
Typical examples for models are:

- the risk prediction for common diseases
- disease course and therapy response prediction
- pharmacokinetic/-dynamic modelling and in silico trial simulations
- data-driven artificial intelligence (AI) models

Standards for Models

For computer models of biological / biomedical systems a selection of the following standards should be used.
They are often defined by community efforts like the COMBINE[107] consortium and listed in detail in the Annex A of the ISO/TS 9491-1 document "general recommendations and requirements for modelling in personalised medicine" as well as in the Annex A and Annex B of the ISO 20691 document "Biotechnology — Requirements for data formatting and description in the life sciences". They are also registered in different registries like for instance the BioSimulators FAIRsharing collection[108], the Biosimulation website[109] and the Normsys registry[110].
The most important standards in this category are:

- Systems Biology Markup Language (SBML), an XML-based description and exchange format for differential-equation models of biological processes. Level 3 of SBML is a modular format that allows to define extension packages for other than differential-equation models.
- CellML, an XML-based description and exchange format for cellular models.
- Biological Pathways eXchange (BioPAX) for the exchange and visualisation of biological pathway data.
- Synthetic Biology Open Language (SBOL) describing the exchange of synthetic biological genetic parts, devices, modules, and systems.
- Neuroscience eXtensible Markup Language (NeuroML) is an XML-based description and exchange format for models in neuroscience consisting of the 4 parts Biophysics, ChannelML, MorphML, and NetworkML.
- The Pharmacometrics Markup Language (PharmML) is an exchange format for pharmacokinetic and pharmacodynamic models.

A more extensive list of formats for biomedical computer models is given in Table 6 in the annex.
For multicellular models there are currently 3 standards (MorpheusML, MultiCellDS and MultiCellML). One should check if one can unite these three standards into one.

Standards for model simulations and documentation of results

For the simulation of systems biology models one of the standard formats listed in Table 7 in the annex shall be used.

Standards for Graphical Model Visualisation

- KEGG (Kyoto Encyclopaedia of Genes and Genomes) Markup Language (KGML).

[107] https://co.mbine.org
[108] https://fairsharing.org/1536
[109] https://docs.biosimulations.org/
[110] https://normsys.h-its.org

- Systems Biology Graphical Notation (SBGN), a graphical notation for representing biological processes.
- Systems Biology Open Language (SBOL-Visual) is a graphical notation to specify genetic parts, devices, modules, and systems.

6.2.3 Standards for metadata of data and models – semantic annotation and taxonomy

It has been estimated that 30% of the world generated data comes from the healthcare sector. Despite the considerable potential offered by all this data, offered by all this data, for many reasons cannot be properly access and re-used. The main reason seems to be the lack of a standardized approach to represent such information. This poses several challenges:

- Dataset coming from different sources cannot be combined
- The meaning of variables contained in the dataset might be not understandable
- Unstructured data cannot be efficiently used to feed models, as, beside the variable per se, other pieces of information might be lacking, *e.g.* unit of measurements;
- Variables within a dataset are usually not systematically related one to each other, losing in this way valuable information along the way.

The **use of a common basis for data representation** is essential to meet the FAIR principles, in particular the concept of interoperability[111]. Within the VHT there is an exquisite variety in technologies to generate data but also create and solve models, that is representative of the heterogeneity of the questions that need to be answered in human health. This further exacerbates the already challenging situation mentioned above and several sub-communities have started trying to organise themselves (*e.g.* logical modelling[112]). Furthermore, all the above points apply not only for the human intelligence, but also to "machine intelligence"[113]. In fact, applying a standard method for representing data would facilitate the development of automated data processing. One solution is represented by implementing a semantic annotation procedure fully tailored to the needs of the VHT.

Semantic annotation is the process to ascribe a specific meaning to elements in a dataset or following a structured and standardized manner, usually based on ontologies, being collections of terms, relational expressions, and associated natural-language definitions designed to capture the intended interpretation of these definitions (Figure 6). Ontologies are necessary when it comes to multi-scale representations, since they can provide a taxonomy and, thus, specifying how the variables are related one to each other. There are two ISO documents[114,115], which recommend the **use of ontologies and standard terminologies** for the descriptions of entities and concepts. They also provide recommendations on how to maintain an ontology (*i.e.*, versions control) and the minimum information to report.

[111] Guizzardi G, Data Intell.(2020);2(1–2): 181–191. https://www.doi:org/10.1162/dint_a_00040.
[112] Touré V *et al*., Briefings in Bioinfo (2021); 22(4): bbaa390. https://doi.org/10.1093/bib/bbaa390
[113] Schneider T. and Šimkus M. KI - Künstliche Intelligenz (2020); 34(3):329–353. https://www.doi.org/10.1007/s13218-020-00686-3.
[114] ISO 20691:2022 for "Requirements for data formatting and description in the life sciences"
[115] ISO/TS 9491-1 for "Recommendations and requirements for predictive computational models in personalized medicine research"

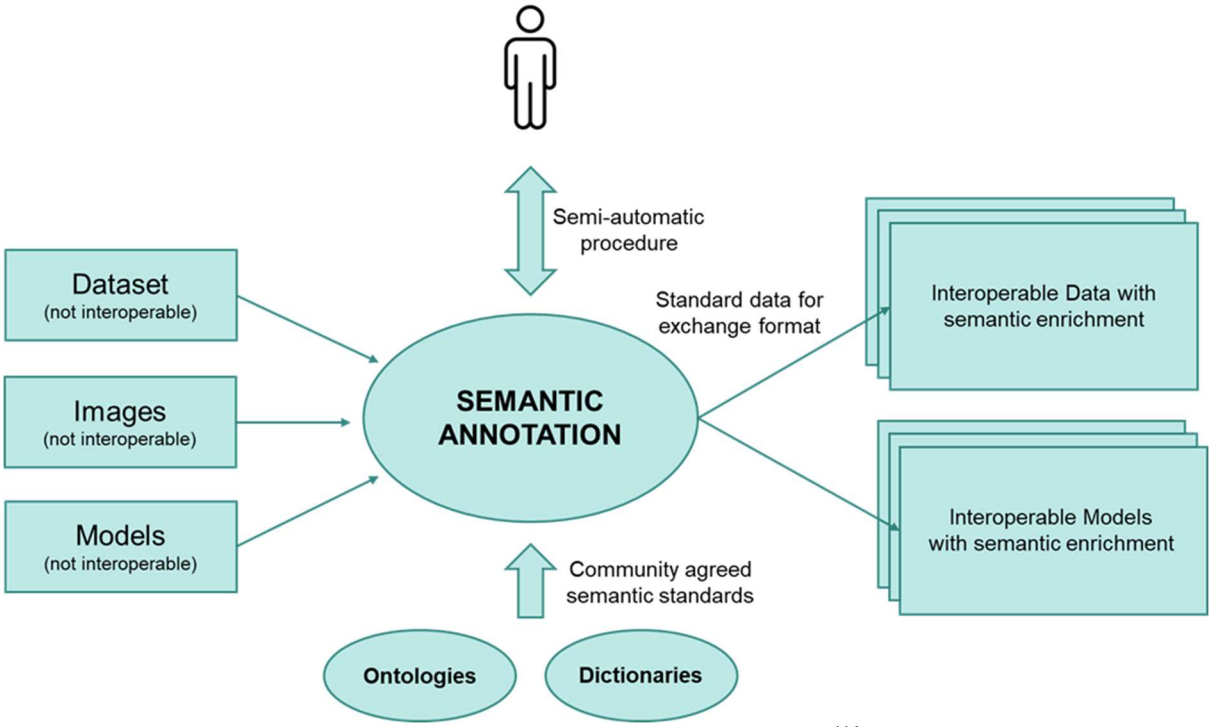

Figure 6: Semantic annotation. Adapted from[116].

Metadata requirements

Annotation of DTHs and their parts, components and entities with metadata is a prerequisite to make the models and their comprised entities **FAIR**. For integrating data from heterogeneous sources the often different naming and coding conventions of metadata attributes must be taken into account. Examples of heterogeneous sources with different conventions are systems like **FHIR** (Fast Healthcare Interoperability Resources), **openEHR** (open Electronic Health Records), **OMOP** (Observational Medical Outcome Partnership) and **CDISC** (Clinical Data Interchange Standards Consortium). For mapping all metadata attributes from all source formats, an intermediate convergence format ("metadata crosswalk"), which includes the maximum set of metadata attributes, can be used[117]. Another challenge is that data fields are often populated with values of different data types. Sometimes data fields from the input format must be split or merged to achieve a mapping to the target format.

The **ISO 20691:2022** standard "*Requirements for data formatting and description in the life sciences*" provides recommendations and requirements for the semantic description and annotation of data in the life sciences (including construction and validation data for computational models and DTHs). This includes a set of minimum consensus information for the annotation of biological data. Metadata and the annotation of the data integrated into DHTs should also consider domain-specific minimum information guidelines. A complete list of minimum information guidelines is given in Annex B.2 of ISO 20691:2022 "*Biotechnology — Requirements for data formatting and description in the life sciences*".

Specifically for predictive computational models in personalised medicine the ISO standard **ISO/TS 9491-1** "Biotechnology — Recommendations and requirements for predictive computational models in personalised medicine research — Part 1: Guidelines for constructing, verifying and validating models" comprises recommendations and requirements for the semantic description and annotation of the models and data that is integrated into them.

[116] Sasse J *et al*. Appl. Sci(2022); 12(2):796. https://www.doi.org/10.3390/app12020796.
[117] Bönisch *et al*., 2022

For systems biology models **MIRIAM** (Minimum Information Requested in the Annotation of biochemical Models) and for simulations **MIASE** (Minimum Information About a Simulation Experiment) could be relevant. **MINIMAR** (Minimum Information for Medical AI Reporting) is a proposed standard for reporting of healthcare prediction models based on artificial intelligence methods.

A minimum information guideline for annotating parameter identifiability and model inference – something like a Minimal Information for Model Inference and Parametrization (**MIMIP**) - is still missing[118]. **PETab** is a format for specifying parameter estimation problems in systems biology in a tabular format.

The metadata attributes and entities should be described by using domain-specific standard terminologies, controlled vocabularies and ontologies as defined in the Annex B.3 of ISO 20691:2022. Such terms describe for instance the species, sex, age, organ, tissue, cell type, disease, and so on. Annotations for identifiable objects or processes, manipulated entities or used technologies are also possible. Typical metadata for modelling and simulations are terms for the used modelling parameters and the obtained simulation results.

For the exchange of metadata information, one or several of the following metadata formats should be used if feasible:

- **ISA-Tab**: **Investigation-Study-Assay**, a Tab-separated format with data, model and Standard Operating Procedures (SOP's) stored under an ISA instance
- **ISA-JSON**: Investigation-Study-Assay in JSON format
- **FHIR**: Fast Healthcare Interoperability Resources, a collection of semantically richly annotated healthcare data

Minimum reporting guidelines for Clinical Decision Support Systems (CDSS)

Part 2 of ISO/TS 9491 describes CDSSs based on AI models. MINIMAR (Minimum Information for Medical AI Reporting) describes the minimum set of required metadata. It also describes recommendations for clinical studies involving computational models for personalised medicine.
Other minimum reporting guidelines for a CDSS, are defined by the EQUATOR[119] (Enhancing the QUAlity and Transparency Of health Research) network or by the TRIPOD[120] (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) consortium.
Examples are CONSORT (Consolidated Standards of Reporting Trials) for reporting of randomised trials, SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials), a standard for trial protocols, STARD (STAndards for the Reporting of Diagnostic accuracy studies), DECIDE (Developmental and Exploratory Clinical Investigations of DEcision), CLAIM (Checklist for Artificial Intelligence in Medical Imaging), PRISMA (Preferred Reporting Items for Systematic Review and Meta-analysis) and PROBAST (Prediction model Risk Of Bias Assessment Tool).
There is a corresponding set of guidelines for models and trials, which are based on Artificial Intelligence (AI) techniques. These guidelines have some AI-specific items added to the checklist, *e.g.*

- TRIPOD-AI: for diagnostic or prognostic prediction models based on AI.
- CONSORT-AI: for clinical trials evaluating interventions with an AI component.
- SPIRIT-AI: for clinical trial protocols with an AI component.
- STARD-AI: for reporting of diagnostic accuracy studies based on AI.
- DECIDE-AI: for CDSSs driven by AI.

Part 2 of ISO/TS 9491 also describes the requirements and recommendations for SOP's (Standard Operating Procedures), the data collection by using the Clinical Record Forms (CRFs), the data cleaning, pre-processing and post-processing steps to ensure the readiness of the data for use by the computational models and the data analytics pipeline with the steps study design, pre-processing, feature selection, model training and validation. Furthermore, it describes the design of integrated data repositories and the data modelling process.

[118] Hucka *et al*. 2015
[119] https://www.equator-network.org
[120] https://www.tripod-statement.org

Terminologies and ontologies for the description and annotation of data, models and their components

Ontologies for Systems Biology / Systems Medicine / Systems Pharmacology simulation shall be used for the semantic annotation of algorithms and the dynamic behaviour of models.
Some ontologies are listed in the Annex B.3 of the ISO 20691[121] document. Others are linked by ontology web portals like *e.g.*, the Open Lookup Service[122] (OLS), BioPortal[123], OboFoundry[124] OntoBee[125] or the semantic lookup platform[126].

Ontologies typically used for in silico medicine modelling and simulation are (a more comprehensive list of ontologies is given in Table 8 in the annex):

- Computational Neuroscience Ontology (CNO).
- An ontology to characterise differences in versions of computational biology Models (COMODI).
- Human Physiology Simulation Ontology containing concepts for simulations, models and algorithms (HUPSON).
- Kinetic Simulation Algorithm Ontology for describing simulation algorithms (KiSAO).
- Mathematical Modelling Ontology (MAMO).
- Ontology of Physics for Biology (OPB) containing physics concepts to describe the dynamics of biological systems.
- An ontology for pharmacokinetic models (PK ontology).
- Precision Medicine Ontology (PreMedOnto[127]).
- Systems Biology Ontology (SBO) with terms useful for describing computational modelling.
- Terminology for the Description of Dynamics (TEDDY), an Ontology for the description of control elements and dynamics in systems biology and synthetic biology.
- Unified Code for Units of Measure (UCUM), defined by the Regenstrief institute.

OpenMinds (open Metadata Initiative for Neuroscience Data Structures) is a metadata framework for neuroscience graph databases defined by the Human Brain (HBP) and EBRAINS projects. It contains metadata models, libraries of controlled terminologies, brain atlases and common coordinate spaces for neuroscience graph databases. Each such metadata model consists of modular metadata schemas defining descriptions of entities from neuroscience.

Clinical languages, terminologies and code systems

Clinical terminologies and code systems are mainly used for semantic annotation of Electronic Health Record (EHR) data and in FHIR[128] resources. The clinical languages like Arden syntax and CQL allow the representation of medical knowledge in Clinical Decision Support Systems (CDSSs) and can be used as a basis for explainability components.

The main clinical languages and terminologies are the following (a more comprehensive list of clinical languages, terminologies, and code systems is given in Table 9 in the annex; further medical vocabularies are also listed in the Unified Medical Language System (UMLS)[129]):

- International Classification of Diseases (ICD-11[130])
- Logical Observation Identifiers Names and Codes (LOINC[131]) for reporting laboratory test results.
- Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT[132]).

Semantic annotation & Taxonomy for VHT

Metadata will be implemented for the uplift of the information coming from the data. Each data object will be accompanied by three sets of metadata:

[121] https://www.iso.org/standard/68848.html
[122] https://www.ebi.ac.uk/ols/index
[123] https://bioportal.bioontology.org
[124] https://obofoundry.org
[125] https://ontobee.org
[126] https://semanticlookup.zbmed.de
[127] https://bioportal.bioontology.org/ontologies/PREMEDONTO
[128] https://www.hl7.org/fhir/resourcelist.html
[129] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html
[130] https://icd.org
[131] https://loinc.org
[132] https://www.snomed.org

- Organizational paradigm metadata – they are associated to the six dimensions to describe the data object (both DOT and DOP). Such metadata are essential for the uplift of the information for that type of data
- Functional metadata – minimum metadata for the VHT being actionable
- Accessibility metadata – they will give information about the "appropriateness of use" from different point of views:
    - Legal – conditions for which the person from which that data has been gathered gave his or her consent
    - Ethical – approval number from the ethical committee
    - Terms of use – terms under which the resource can be used ("civil code of supplementary laws") given by the user who upload the resource
    - Accessibility – how the data object can be accessed
    - Credibility – information about trueness and precision
- Optional metadata – metadata that could be added to better describe the data object, but not essential for the implementation of the VHT (*i.e.*, atlas positioning through the six dimensions and actionability)

There are also two metadata transversal categories:

- Necessarily standardized metadata - Metadata to make the infrastructure work, for which the content must necessarily be standardized
- Optional standardized metadata - Metadata for which a precise standard is not necessary

The metadata belonging to the first category are those essential for processes automatization, essential either for the infrastructure or for describing the data object. The minimum set of annotations will be decided by the community. Semantic mediation methods will be implemented to foster the upload of data objects also when coded with different standards. However, since the semantic mediation do not allow for a perfect mapping from a standard to another one, a human control will be added at the end of the procedure to check the quality of the mapping. This applies for publication purposes on the catalogue. If the data object has to be used in the platform for simulation, it must be coded with the standard adopted within the VHT (chosen by the community). Another possible approach is the Common Data Model – the header is pre-defined, the user can then decide how to code the information.

6.3 Health Technology Assessment and Payers

Health technology assessment (HTA) is a multidisciplinary process that uses systematic and explicit methods to evaluate the properties and effects of a health technology[133]. It can be used for multiple purposes, but the most common is to evaluate the cost-effectiveness of new treatments of instrumentation, and to support reimbursability evaluations.

So far, to the author's knowledge, no concrete use of in silico methodologies has been reported to support HTA decisions. On the other hand, a few HTA assessment of Digital Twin technologies are starting to appear[134]. The HTA assessment of a digital twin does not pose particular challenges; the HTA framework is general enough to be used for any technology, whether predictive or not. More interesting is the use in silico methodologies for the HTA assessment of new medical products, whether drugs or devices. The use of computer modelling and simulation in Health Technology Assessment could offer several advantages. In silico methods can be a cost-effective and time- efficient approach compared to conducting real-world trials or traditional clinical experiments. Also, computer models can estimate parameters that are difficult or impossible to measure and can predict long-term outcomes, enabling the assessment of potential impacts and benefits of new medical products over extended periods.

While the use of in silico methodologies seems promising for improving HTA processes, it is essential to examine its limitations and potential risks, especially in the context of Responsible Research & Innovation related to the use of the virtual human. Considering the barriers that need to be addressed for the VHT to realise its full potential in advancing human health (Section 4.2), the ones that more closely interests the application on HTA and Payers are mostly related to data availability and quality, lack of standardization, technical expertise and resources, validation and transparency:

- **Data availability and quality**: The availability of high-quality data is crucial to assess the validity and reliability of the modelling results.

[133] https://en.wikipedia.org/wiki/Health_technology_assessment
[134] https://www.nice.org.uk/guidance/mtg32

- **Lack of standardization**: the lack of standardized methods, guidelines, or best practices to demonstrate the efficacy and cost-effectiveness of in silico methodologies in HTA can lead to inconsistencies in approaches, difficulties in interpreting and evaluating different models and results.
- **Technical expertise and resources**: Payers and HTA bodies may face barriers in terms of expertise, infrastructure, and funding necessary to utilize these models effectively. Also, introducing new methodologies and approaches, can face resistance due to cultural barriers and lack of familiarity especially when making time-sensitive decisions, which can limit the feasibility and practicality of using complex models.
- **Validation and transparency**: Validating and transparently documenting computer models and simulations is essential to assess their credibility. However, validating complex models can be challenging, and lack of transparency in model assumptions, algorithms, and data sources can hinder trust and acceptance among stakeholders, including payers.

The VHT can surely support and facilitate these research activities aimed to explore the usefulness and the limits of using in silico methodologies for the HTA of medical products. but in addition to that dedicated research funding should be devolved to this.

6.4 Legal aspects

The legal analyses carried out here started from 7 macro-questions:

1. which are the privacy conditions and safeguards that must be fulfilled so that personal data can be lawfully used in the context of VHT?;
2. is the reuse of health data permitted in the EU for the purpose of delivering Artificial Intelligence-driven medical solutions?
3. which obligations apply to the developers and the users of AI-based models necessary to elaborate VHT and relevant ecosystems?
4. how to better connect advanced models to emerging data generation technologies (such as IoT and Edge AI wearables, implantables and point of care sensors) provided their ability to capture fast disease dynamics and enable prevention and personalisation in clinical settings?
5. can specific Privacy-Enhancing Technologies help ensure safe and compliant processing for the purpose of in silico medicine?
6. can clinically reliable VHT be generated thanks to anonymous, pseudonymous or synthetic data?
7. which regulatory recommendations can be made to policy-makers to ensure that the EU will soon be at the global forefront of the VHT sector?

We then identified 5 main areas of legislation which, to date, are relevant to In Silico Medicine and Virtual Human Twins: a) **Privacy and Data Protection** (Regulation (EU) 2016/679, GDPR' and national associated laws); b) **Data Governance and reuse** (to date, the GDPR and, near future, the Data Governance Act, the Data Act and the European Health Data Space); c) **Clinical Trials (Regulation** no. 2014/536 and related legislation, such as the Good Clinical Practice Directive); d) **Medical Devices** (Regulation no. 2017/745 and the In Vitro Medical Device Regulation no. 2017/746, as well as connected laws such as the Health Technology Assessment Regulation); e) **Artificial Intelligence** (Artificial Intelligence Act).

6.4.1 Data privacy & protection

General Data Protection Regulation

Due to the AI-driven nature of the VHT validation and patient-specific-modelling processes, the development of in-silico applications can only be achieved through the collection, analysis, combination, and use (in a word, 'processing') of information which, in most of the cases, relates to identified or identifiable persons, so falling in the legal definition of 'personal data'. As of 2018, the processing of this kind of information in the European Union is governed by the pioneering rules of Regulation (EU) 2016/679 ('General Data Protection Regulation' or 'GDPR'). This piece of legislation has quickly become the benchmark of data protection framework and standards at international level. Innovative principles (such as accountability, or privacy-by-design and by-default) and significant

obligations (such as transparency and security requirements) are established, in both public and private sectors, for data controllers (namely the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of the personal data) and data processors (*i.e.* the natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller) aimed at guaranteeing the individual right to the protection of personal data. The GDPR gives each person full control over any kind of processing of his/her personal data – especially providing individuals with specific rights that may be exercised vis-à-vis data controllers (and also processors) – and introduced relevant fines in the event of non-compliance.

***Relevance to the VHT*:** The rules laid down by the GDPR have a wide impact on the VHT, considering that the regulation includes specific prescriptions on the processing of health data (which fall under 'special categories of personal data'), processing for scientific research purposes, secondary use of personal data, anonymization and pseudonymisation, automated decision-making, technical and organisational security measures and codes of conduct.

6.4.2 Data governance

Governance has become a crucial aspect in the regulation of data, as they hold an intrinsic economic value. Through the years, the European Legislator has therefore approved a series of norms to improve data sharing among public entities, citizens and private companies, with a view of further developing the borderless digital internal market. The present section aims to offer a brief overview of the major Regulations and Directives related to the governance of personal and non-personal data.

Data governance Act

On 23rd June 2022, the European Union approved the **Data Governance Act (reg. 2022/868)** with the scope of fostering trust in **data sharing** and overcoming technical difficulties involving the **reuse of data**[135]. This Regulation is part of the **European Strategy for Data**[136], as defined by the European Commission with the aim of improving the digital internal market. The European legislator recognized that a harmonised framework for data exchanges could lead to the improvement of strategic sectors, such as health, mobility, energy, and agriculture[137]. For this reason, the Regulation, entering into force by September 2023[138] defines a legal framework for the governance of the reuse of data that are held by public entities and data subjects.

Concerning the governance of data held by public sector bodies, the Data Governance Act defines a list of **data**[139], **protected** for reasons of IP rights, data protection, commercial confidentiality, and statistical confidentiality, that public entities can share with people interested in their reuse for scopes of public interest (data users). Public entities are free to decide the **conditions** set for reuse and make them public[140]. To use personal data, data users will need to demonstrate a **legitimate basis** according to GDPR for the processing[141]. This means that Regulation 2022/868 does not derogate from the provision set in Regulation 2016/679, being both simultaneously applicable[142].

The whole process of reuse is characterised by **transparency**: the Member State designates one or more competent bodies, which assist public sector bodies in the governance of access to data while providing guidance and technical support on how to best structure and store data[143].

[135] Rec. 3, DGA; Art. 1, DGA.
[136] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data, 19.2.202, COM(2020)66 final.
[137] Rec. 2, DGA.
[138] Art. 38, DGA.
[139] Art. 3, DGA.
[140] Art. 5 (1-3), DGA
[141] Art. 5 (6), DGA.
[142] Rec. 7, DGA.
[143] Art. 7, DGA.

Data Governance Act aims at improving sharing of data between private data holders and potential data users[144]. In this regard, new actors are defined, such as **data intermediation services providers**[145] and **data altruism organisations**[146]. Both have specific conditions to comply with, and a notification regime for being recognized in all Member States.
On one side, **data altruism** is characterised by the voluntary sharing of data on the basis of the consent of data subjects for objectives of **general interest**, implying that the reuse of data is to be guaranteed without lucrative scopes[147]. General interest can involve scientific research, combating climate change of improving public policies. A European data altruism consent form is defined by the Commission to facilitate the collection of personal data based on data altruism[148].
On the other side, **data intermediation services providers** ensure a series of services aiming at establishing commercial relationships between data subjects and data holders or between data holders and data users[149]. They can pursue lucrative scopes, although they must ensure that the procedure to access to services is compliant with the principles of transparency, fairness and non-discrimination.

Relevance to the VHT: VHT is based on a large amount of personal data, which serves the purpose of training and limiting the error level of this technology. The Data Governance Act in this respect could improve the development of VHTs by facilitating the access to personal data, especially in light of the data altruism structure that the Regulation itself defines.

Data Act

As part of its data strategy (and following the Data Governance Act discussed above), in February 2022 the European Commission launched the Data Act proposal (COM(2022) 68 final). The Data Act proposes new rules on who can use and access data generated in the EU across all economic sectors, with the intent to unlock the economic and societal potential of data and technologies in line with EU rules and values. It imposes obligations on private data holders to make data available to public sector bodies under exceptional circumstances, but also maintains incentives for manufacturers to continue investing in high-quality data generation. In the spring 2023 the Data Act entered the trilogue between Council and Parliament, and a conclusion of negotiations appears possible before 2024.

European Health Data Space

As part of the European Data Strategy, the European Health Data Space (EHDS) is conceived as the first European space aiming at improving the exchange and **sharing of electronic health data** throughout the Union[150], for so-called primary and secondary purposes. The proposed regulation therefore aims to facilitate the processing of health data by strengthening the rights of data subjects, establishing a governance framework for primary and secondary use, and identifying conditions for the development of electronic health record systems[151]. New definitions are introduced, which should be understood to supplement those already present in the GDPR and the DGA. Specifically, **primary use** is defined as the processing of electronic personal health data for the purpose of assessing, maintaining

[144] Rec. 2-3, DGA.
[145] Art. 2 (11), DGA: «data intermediation service' means a service which aims to establish commercial relationships for the purposes of data sharing between an undetermined number of data subjects and data holders on the one hand and data users on the other, through technical, legal or other means, including for the purpose of exercising the rights of data subjects in relation to personal data, excluding at least the following: (a) services that obtain data from data holders and aggregate, enrich or transform the data for the purpose of adding substantial value to it and license the use of the resulting data to data users, without establishing a commercial relationship between data holders and data users; (b) services that focus on the intermediation of copyright-protected content; (c) services that are exclusively used by one data holder in order to enable the use of the data held by that data holder, or that are used by multiple legal persons in a closed group, including supplier or customer relationships or collaborations established by contract, in particular those that have as a main objective to ensure the functionalities of objects and devices connected to the Internet of Things; (d)data sharing services offered by public sector bodies that do not aim to establish commercial relationships».
[146] Chapter IV, DGA.
[147] Art. 2 (16): «'data altruism' means the voluntary sharing of data on the basis of the consent of data subjects to process personal data pertaining to them, or permissions of data holders to allow the use of their non-personal data without seeking or receiving a reward that goes beyond compensation related to the costs that they incur where they make their data available for objectives of general interest as provided for in national law, where applicable, such as healthcare, combating climate change, improving mobility, facilitating the development, production and dissemination of official statistics, improving the provision of public services, public policy making or scientific research purposes in the general interest».
[148] Art. 25, DGA.
[149] Chapter III, DGA.
[150] Rec. 1 EHDS.
[151] Art. 1 (2), EHDS..

or restoring the health status of the natural person to whom it relates[152]. **Secondary use** is defined as the processing purposes solely established in the EHDS, regardless of the source from which the data originate[153]. **Data owner** and **data user** are also defined as, respectively, the one who has the right or obligation to make personal electronic health data available or the ability to share non-personal electronic health data[154], and the one who has the right to access these data[155].

Referring to the provisions of the GDPR, the EHDS will extend the application of data subjects' **rights of access** and portability[156]; with regard to the former, Member States must be able to guarantee immediate, free and easily readable, consolidated and accessible access to the electronic personal health data of data subjects, who always have the right to receive an electronic copy of their data. However, given the ethical and moral implications that could result from unrestricted access, Member States may provide **limitations** about its exercise. With regard to the **right to portability**[157], it is provided that patients may request consultation of their data to the health care provider they designate, without the legal basis provided for the processing being based on the consent of the data subject or the performance of a contract.

Health care professionals can have access to the electronic health data of patients in their care and ensure that it is kept up to date. EHDS establishes the **priority categories of electronic health data**[158] to which each Member State must provide access for primary use and which must necessarily be recorded by each health professional in an electronic health record system.

Regarding **electronic health record** (EHR) systems, the proposal reserves a number of obligations for manufacturers[159], importers[160] and distributors[161] of EHR systems. Similar to reg. 2017/745 (Medical Device Regulation) and reg. 2017/746 (In Vitro Medical Device Regulation), it defines the documentation required for the purpose of placing the EHR system on the market[162], the obligation to appoint an authorised representative if the manufacturer of such a system is established outside the Union[163], the obligation of interoperability of high-risk AI systems with EHR systems[164].

Electronic health data are also a relevant resource for purposes other than disease treatment and prevention. For this reason, the EHDS provides a number of purposes for which the exchange of these data must be ensured[165]. Among the purposes envisioned as secondary uses, **scientific research,** training of AI systems and digital health applications, and development and innovation activities for products or services that contribute to public health come into relevance. Bodies responsible for access, appointed by Member States, ensure that users will only have **access to anonymized or pseudonymized data** if the purpose cannot otherwise be served[166].

[152] Art. 2 (2, d): «'primary use of electronic health data' means the processing of personal electronic health data for the provision of health services to assess, maintain or restore the state of health of the natural person to whom that data relates, including the prescription, dispensation and provision of medicinal products and medical devices, as well as for relevant social security, administrative or reimbursement services;».
[153] Art. 2 (2, e): «'secondary use of electronic health data' means the processing of electronic health data for purposes set out in Chapter IV of this Regulation. The data used may include personal electronic health data initially collected in the context of primary use, but also electronic health data collected for the purpose of the secondary use».
[154] Art. 2 (2 ,y): «'data holder' means any natural or legal person, which is an entity or a body in the health or care sector, or performing research in relation to these sectors, as well as Union institutions, bodies, offices and agencies who has the right or obligation, in accordance with this Regulation, applicable Union law or national legislation implementing Union law, or in the case of non-personal data, through control of the technical design of a product and related services, the ability to make available, including to register, provide, restrict access or exchange certain data».
[155] Art. 2 (2, z): «'data user' means a natural or legal person who has lawful access to personal or non-personal electronic health data for secondary use».
[156] Art. 3 (1), EHDS.
[157] Art. 3 (3), EHDS.
[158] Art. 5, EHDS.
[159] Chapter III, Section 2, EHDS.
[160] Art. 19, EHDS.
[161] Art. 20, EHDS.
[162] Art. 24, EHDS.
[163] Art. 18, EHDS.
[164] Art. 14, EHDS.
[165] Art. 34, EHDS.
[166] Art. 44, EHDS

Relevance to the VHT **:** EHDS is the first step taken by the EU to improve data sharing throughout the territory of the Union. VHT strongly relies on health data as a fundamental source for its development. Therefore, by defining the essential issue of primary and secondary use of health data, EHDS is of pivotal importance for VHT technology.

6.4.3 Artificial Intelligence

VHT is an integrated, multiscale, multi-time and multi-discipline representation of quantitative human physiology and pathology[167]. Artificial Intelligence is therefore a crucial component in the development of VHT for purposes of healthcare and medical scientific research.

The European Union has still not adopted a comprehensive legal framework governing Artificial Intelligence. However, the European Commission has been particularly active in this respect. In 2021, the Communication on Fostering a European Approach to Artificial Intelligence was issued, presenting a legislative strategy for the regulation of AI[168] revolving around a risk-based approach. This Communication translated, later on, into the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence and the Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to Artificial Intelligence.

AI Act

On 21st April 2021, the European Commission issued the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)[169]. The proposal belongs to the European AI Strategy[170], aiming to support research and industrial capacity while protecting fundamental rights. It was delivered by the Commission after the publication of the «White Paper on AI – A European approach to excellence and trust»[171], which set out the policy options on how to achieve both the promotion of the uptake of AI and the protection of fundamental rights from risks associated with the use of such technologies.

The Proposal introduces a classification of Artificial Intelligence systems, divided on the basis of the purpose of use. Three different categories are set: the prohibited AI practices[172]; the high-risk AI Systems[173], and the low-risk AI[174]. The high-risk AI systems must comply with specific requirements before being put on the market. This implies that, depending on the category under which the AI System falls, different obligations apply to the active subjects of the supply chain (producers[175], importers[176], distributors[177], users[178]). Among these, manufacturers must adopt a risk management system for the entire lifecycle of the system[179], maintaining technical documentation[180] and keeping records[181]. They must ensure that the AI system is compliant with transparency[182] and human oversight criteria[183]. Also, the system must achieve an appropriate level of cybersecurity and robustness.

The Proposal introduces the "Notified Bodies", which must ensure the conformity of high-risk AI systems to the provisions set by the AI Act[184]. They are part of a broader scheme of allocation of risks,

[167] European Virtual Human Twin (edith-csa.eu)
[168] Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions. Fostering a European Approach to Artificial Intelligence' COM(2021) 205 final
[169] Proposal for a Regulation of the European Parliament and the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, COM(2021) 206 final.
[170] A European approach to artificial intelligence | Shaping Europe's digital future (europa.eu)
[171] White Paper on Artificial Intelligence – A European approach to excellence and trust, Brussels, COM(2020)65, 19.2.2020
[172] Art. 5, AI ACT.
[173] Title III, AI ACT.
[174] Title IV, AI ACT.
[175] Art. 16 -25, AI ACT.
[176] Art. 26, AI ACT.
[177] Art. 27, AI ACT.
[178] Art. 28-29, AI ACT.
[179] Art. 9, AI ACT.
[180] Art. 18, AI ACT.
[181] Art. 20, AI ACT.
[182] Art. 12, AI ACT.
[183] Art. 14, AI ACT.
[184] Art. 33, AI ACT.

that is also based on a regime of serious incident reporting to national authorities[185] and possible corrective actions[186].

AI Liability Directive

The proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence[187] is part of the European AI Strategy. Alongside with the Proposal for a regulation known as 'AI Act', which deals with the framing of different enforcement regimes depending on the level of risk of artificial intelligence, the Directive aims to set general conditions for the imputability of damages arising from artificial intelligence.

Over the years, there have been many proposals regarding damages. Strict liability, concerning the developers of artificial intelligence, has been envisaged on the basis of liability arising from dangerous activities. However, the Proposal settles on the imputability of the damage to the provider or user on the basis of risk management. In this respect, it establishes the conditions under which the causal link[188] between the fault of the defendant in a damages action and the output or failure to output produced by the IA system is presumed to exist. Specifically, the claimant has to demonstrate specific elements provided by the Proposal, e. g. the non-compliance with a duty of care provided by the AI Act from the defendant[189] and causal link between the output produced by the AI system and the damage[190].

***Relevance to the VHT*:** The liability regime outlined in the AI Liability Directive clarifies and completes the framework of entities responsible for the manufacture, distribution, and use of AI. It therefore represents a useful and fundamental tool for the purposes of the development and distribution of AI-devices for medical purposes, VHT included.

6.4.4 Medical devices

Throughout the years, the European Union has been proactive in the regulation of the putting in the market of medical devices. In this respect, it is worth noting that the first legal framework consisted of **three Directives** (Council Directive 90/385/EEC on Active Implantable Medical Devices, Council Directive 93/42/EEC on Medical Devices, and Council Directive 98/79/EC on In Vitro Diagnostic Medical Devices), which only in recent years were reformed by the European Legislator. Today, the legal framework governing the development and the putting in the market of medical devices is formed by the **Medical Device Regulation** (reg. 2017/744) and the **In Vitro Medical Device Regulation** (2017/746).

In this regard, the two regulations are intended to enforce already existing elements, such as the supervision of notified bodies, conformity assessment procedures, clinical investigations and clinical evaluation, vigilance and market surveillance[191], while introducing new ones. In this regard, it is to be noted that the material scope of application of MDR and IVDR is broader, the former involving certain aesthetic devices[192] and software as well, and the latter being applicable also to diagnostic services[193].

At the EU level, the **Medical Device Coordination Group**[194] is designated to provide guidelines and orientations of the provisions of the MDR and IVDR to the Commission and Member States. The Medical Device Coordination Group is formed by persons designated by the Member States based on their role and expertise in the field of medical devices. National competent authorities can cooperate among each other and the Coordination Group to provide additional support[195].

It is relevant to note that, while falling under the scope of application of the Medical Device Regulation, VHT technologies may not be regulated by the In Vitro Medical Device Regulation since they are not intended to be directly used in the examination of specimens derived from the human body.

[185] Art. 62, AI ACT.
[186] Chapter III, AI ACT.
[187] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final.
[188] Art. 4, AILD.
[189] Art. 4, p. 1 (a), AILD.
[190] Art. 4, p. 1 (c), AILD.
[191] Rec. 4, MDR.
[192] Rec. 12, MDR.
[193] Art. 2, IVDR.
[194] Art. 103, MDR.
[195] Art. 102, MDR.

Medical Device Regulation

Medical Device Regulation

In May 2017 the European Union approved the **Medical Device Regulation (reg. 2017/745)**, which entered into force on 26th May 2021. The Regulation introduces new **requirements for the manufacturers** to be placed on the market, making available on the market or putting into service medical devices and related accessories for human use. It replaced the former Directive on Implantable medical devices and the Medical Device Directive.

It introduces a new definition of Medical Devices[196], including software, and regulates their place on the market according to a risk-based approach. It introduces a **classification of devices**, divided according to the intended purpose of their functioning and their inherent risks[197]. In particular, the Regulation set a series of obligations to economic operators concerning – among others - the **technical documentation** to be published[198], and the establishment of a system for **risk management**[199] and **quality management**[200], that may vary depending on the class of the Medical Device. The MDR disciplines also the procedure for the **identification and traceability of devices**[201], and their registration in EUDAMED[202]. Member States can designate a conformity assessment body as a notified body to carry out conformity assessment activities under MDR[203] and supervise the application of the Regulation itself.

Relevance to the VHT: Although it is still discussed whether Artificial Intelligence could be included in the material scope of application[204], the Medical Device Regulation contributes to defining the legal framework applicable to high-tech tools and related accessories. In this respect, the placing into the market of hardware and software components of VHT could be subjected to such norms.

In Vitro Medical Device Regulation

Regulation 2017/746, which entered into force on 26th May 2022 and is gradually replacing all the conditions set by the previous Directive 98/79/EC, introduces a **new definition of In vitro Medical Devices[205]**. Like the MDR, it includes software, which, alone or in combination, is set to operate with specimens derived from the human body. Similarly to MDR, it defines the rules concerning the putting in the market[206] and the obligations falling upon the manufacturers[207] that may depend on the risk class of the In Vitro Device[208]. For example, Classes from B to D, representing the highest levels of risk[209], require an **assessment** and **certification** by a notified body for medical devices prior to being placed

[196] Art. 2 (1) MDR: «'medical device' means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes: diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease; diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability; investigation, replacement or modification of the anatomy or of a physiological or pathological process or state; providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations, and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means»

[197] Art. 51 MDR.

[198] Art. 10 (8) MDR.

[199] Art. 10(2) MDR.

[200] Art. 10 (3) MDR.

[201] Art. 25; Art. 27 MDR.

[202] Art. 30, MDR.

[203] Art. 35 MDR

[204] E. Niemiec, Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices?, in Digital Health, 8, 2022, 2.

[205] Art. 2 (2): «'in vitro diagnostic medical device' means any medical device which is a reagent, reagent product, calibrator, control material, kit, instrument, apparatus, piece of equipment, software or system, whether used alone or in combination, intended by the manufacturer to be used in vitro for the examination of specimens, including blood and tissue donations, derived from the human body, solely or principally for the purpose of providing information on one or more of the following: (a) concerning a physiological or pathological process or state; (b) concerning congenital physical or mental impairments; (c) concerning the predisposition to a medical condition or a disease; (d) to determine the safety and compatibility with potential recipients; (e) to predict treatment response or reactions; (f) to define or monitoring therapeutic measures. Specimen receptacles shall also be deemed to be in vitro diagnostic medical devices;»

[206] More specifically, Chapter II, IVMDR.

[207] Art. 10, IVMDR

[208] Annex IV, IVMDR.

[209] Risk is defined under Art. 2 (16) IVMDR as: «the combination of the probability of occurrence of harm and the severity of that harm»

on the market. As for general medical devices, manufacturers need to ensure the safety and performance requirements[210] as set by IVMD Regulation, as well as the post-market surveillance[211] and vigilance[212].

6.4.5 Remarks and recommendations based on current EU policies

The discussion above focused on 5 main areas of legislation which, to date, are relevant to In Silico Medicine and Virtual Human Twins: Privacy and Data Protection, Data Governance and reuse , Clinical Trials, Medical Devices, Artificial Intelligence.

Focusing on **health data reuse**, it must be highlighted how the intention of the EU legislator to foster scientific research by establishing, in the GDPR, simplified procedures and lighter requirements in connection with the re-use of data (and reference was made particularly to Art. 5.1, b) GDPR, which lays down a presumption of compatibility of the reuse of personal data for scientific research purposes, provided that data minimization measures are adopted such as pseudonymisation), has been in many cases frustrated at national level. Member States were in fact empowered to "introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health" (Art. 9.4 GDPR).
This led to a loss of homogeneity in the way the EU legislation on the circulation and protection of health data was integrated at local level, "resulting in a complex and fragmented landscape for researchers to navigate. Consequently, differences between Member States in the way the GDPR is implemented and interpreted in the area of scientific research has made data exchange between Member State and EU bodies for research purposes difficult and in some cases highly technical" (Par. 5.1.2 of the 'Assessment of the EU Member States' rules on health data in the light of GDPR' by the EU Commission's Health and Food Safety Directorate-General). In brief, sharing and re-using health data for medical and scientific research, both at national and cross-border level, poses to date serious risks of non-compliance. This stokes fear of violations and sanctions and thus concretely prevents scientific progress, as well as the implementation of the European Health Data Space (and, in the end, the creation of the coveted European Research Area).
After having assessed in-depth this fragmentation, the main novelties of the Data Governance Act, the Data Act, the Artificial Intelligence Act and the European Health Data Space were analysed, with a special focus on the roles and responsibilities of the parties involved in the reuse and/or sharing of personal and health data.
In this respect, special attention was paid to **anonymisation and pseudonymisation**, comparing the very complex scenario and operational uncertainties deriving from the strict interpretation of these concepts provided by the Article 29 Working Party (now renamed as European Data Protection Board) in its – still fully applicable – Opinion 05/2014 on Anonymisation Techniques, and the novel perspective offered by EU Court of Justice (CJEU) in a very recent landmark ruling (dated 26 April 2023, in Case T-557/20).
In brief the CJEU highlighted that in order to determine whether pseudonymised information transmitted to a data recipient constitutes personal data, it is necessary to consider the latter's standpoint: if the data recipient does not have any additional information enabling it to re-identify the data subjects and has no legal means available to access such information, the transmitted data can be considered anonymized and, therefore, not personal data. The fact that the controller transmitting the data still has the means to re-identify the individuals is irrelevant and does not mean that the transmitted data is automatically also personal data for the recipient.

Taking into account the above, we concentrated on the EHDS to emphasize that:

- **Data sharing by Data holder**: as clearly stated in Recital 37, the EHDS establishes an obligation for data holders to make the electronic health data they have collected available to data access bodies (Art. 6.1, c) GDPR) and provides for the conditions to derogate the prohibition to process health data in accordance with Articles 9(2) (h),(i),(j) GDPR;

[210] Capter VI, IVMDR.
[211] Chapter VII, Section 1, IVMDR.
[212] Chapter VII, Section 2, IVMDR.

- **Data permit issued by Data access bodies**: Art. 33.5 EHDS states that «Where the consent of the natural person is required by national law, health data access bodies shall rely on the obligations laid down in this Chapter to provide access to electronic health data»;
- **Data access by Data user**: Art. 45 reads that data applicants must demonstrate that both an appropriate legal ground (Art. 6 GDPR) and a valid condition exist (Art. 9.2 GDPR) which permit them to lawfully access and process the data for one or more of the admitted purposes.

Notwithstanding this valuable attempt to attain simplification for health data reuse, the EDPS and EDPB have questioned how the above «may be reconciled with Article 9(4) GDPR and the possibility for Member State law to introduce further conditions, including limitations with regard to the processing of genetic data, biometric data or data concerning health» (Par. 89 of the EDPB-EDPS 'Joint Opinion 03/2022 on the Proposal for a Regulation on the European Health Data Space' adopted on 12 July 2022). In light of all the above considerations, it must be concluded that in order to foster clinical research, in silico medicine and so the uptake of VHT in the EU, the whole triangulation (*i.e.* all data flows) between (i) Data Holders, (ii) Data Access Bodies and (iii) Data Users, should rely on the legal framework established by the EHDS – also in connection with Data Act – by explicit derogation of Art. 9.4 GDPR, meaning that any eventual limitation or stricter condition for accessing and reusing health data laid down by member States at national level must not apply, being overridden by the obligations posed by the EHDS (and the Data Act).

6.4.6 IPR management

The VHT is realised through a collaborative distributed knowledge and resource platform, aiming at improving the adoption of patient-specific predictive computer models[213]. In this sense, the VHT can be considered **software**, as such subject to the application of Dir. 2009/24/EC ("On the legal protection of computer programs")[214]. This piece of legislation aims to remove differences in terms of copyright of software among Member States[215], by according the same protection as literary works[216]. Therefore, nowadays both economic and moral rights are ensured by the law to the creator of the computer program.

On one side, the **moral rights**, *id est* the right to be recognized as the owner of the work and to oppose publication and modifications, are inalienable and imprescriptible, being them recognized as fundamental rights. On the other, **the right to economically exploit intellectual property rights** is protected for **70 years** after the owner's death. Such rights are always alienable[217].

In this respect, it is important to note that the right to economically exploit the work is different from the right to reproduce it. The former is defined as the consequence of recognizing software as an intellectual work, while the latter is the consequence of recognizing the software as a copy. Both can be the object of a contract. However, the right to economically exploit the work is usually sold, while the right to reproduce is generally put into licence.

Considering that VHT is proposed to be implemented in hospitals and healthcare facilities, it is likely that it will be the object of a licensing contract. In this respect, it is important to note that licenses can be different, and terms may change according to the will of the parties involved. However, there are various elements generally present in these acts. For example, the right of sub-licensing the program is usually excluded.

Some licenses can be "**open source**", namely licences where the licensor gives more rights to the licensee than the mere use. The licensee, at the same time, respects certain obligations: for example, the obligation to redistribute the program with the same type of licence. One of the most famous in this regard is the GPL (*Global public licence*), which is based on direct access to the source code of software and to its free modifiability.

Another type of licence is the so-called "freeware", where the contract of licence does not provide for a payment as the equivalent of the right to use the program.

[213] European Virtual Human Twin (edith-csa.eu)
[214] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs.
[215] Rec. 5 Dir. 2009/24/EC.
[216] Rec. 6 Dir. 2009/24/EC.
[217] G. Finocchiaro, *Diritto di internet,* Bologna, 2020, 133.

6.5 Ethical and social aspects

Individuals' decision-making and the functioning of technology guided by personal values can be influenced by ethical principles, which serve as a framework of moral standards. Ethical obligations can provide support and motivation to stakeholders, ensuring the safeguarding and advancement of human values. While legal frameworks establish the formal prerequisites for individuals and organisations, ethical obligations can act as catalysts for what is stipulated by the law. Moral considerations aid in shaping people's actions, imposing limits, and providing guidance for technological advancements. Moreover, ethics can form the foundation of laws, with legislation often originating from ethical quandaries. For instance, concepts like informed consent, privacy, and confidentiality may be intertwined with the principle of individual autonomy. Similarly, anti-discrimination laws can be regarded as stemming from the principle of justice. However, no matter how much the law endeavours to uphold ethical concerns, it may not always prevent morally undesirable consequences from occurring.

Certain technologies have **sparked ethical debates** and **necessitated significant policy initiatives** aimed at controlling their trajectory, typically only after some harm has been caused. By identifying and examining potential concerns, ethics can serve as a tool to assess the risks or advantages associated with tangible forms of harm that may arise from emerging technologies. This is particularly relevant in the case of Artificial Intelligence (AI) and its regulation. Ethical examinations of AI have highlighted potential issues that may require regulation, such as bias and opacity within AI systems. To illustrate, this prompted the EU legislator to outline specific requirements regarding AI in the proposed AI Act. In summary, ethics can complement legal frameworks and contribute to the comprehension of how laws should be applied in novel or unfamiliar circumstances.

Social elements will be further elaborated.

6.6 Conclusions

TBD

7 Uptake of the virtual human twin

7.1 Introduction

TBD

7.2 Users

Recognizing the diversity and complexity inherent in the stakeholders backgrounds, experience and intended use, we understand the importance in well-defined user profiles and roles will play in system efficiency, data security, regulatory compliance, and user experience.

7.2.1 User roles

In the VHT, a **user profile is defined as a collection of settings and information associated with a specific user**. It contains critical information that is used to identify an individual, such as their name, age and individual characteristics such as affiliations, background knowledge or expertise. The profile should be linked to an external IDP (Identity provider) and thus the authentication should be performed with trusted third party services (such as trusted National Identity Provider services in current EU states or an European wide citizens identification system such as the personal digital wallet for EU citizens and residents thanks to the trust framework created by the eIDAS Regulation in the long term.) **User profile does NOT distinguish the role of the user in the system**. User profiles are adaptable as the individual changes jobs, relocating to other countries, acquiring other knowledge or expertise.
User role is a well-defined collection of capabilities within the system allowing the user to achieve the role's intended purposes. In other words, user roles are clusters of system privileges that are designed to achieve specific goals using the VHT infrastructure. User roles are dynamic as their assignment to user profiles. *A user profile can consist of various user roles as a person can have different aims in using the VHT platform in her/his capacity*. New roles can be defined as new paradigms of using the VHT platform emerges. In the initial phase we envision four categories of roles in the VHT.

Healthcare Professional Category (Healthcare professionals, Doctors, Specialists, *etc.*)
Clinical users intend to use the VHT to improve health care supply for individual patients. In order to achieve the integration of VHT-solutions into the clinical workflows, both clinical IT management as well as clinicians have to cooperate, both characterized by specific profiles and roles in the context of the VHT:

- Clinical IT manager:
  - Profile: management of in-house clinical data (repositories, standardization, curation), management of external access to in-house clinical data (management of access to data, data privacy), management of in-house clinical software ecosystem (*e.g.* safety issues), management of hardware resources
  - Roles in context of VHT: interface between clinicians and clinical researchers and the VHT platform in the context of data utilization, running the VHT, installation of software and hardware interface to the clinicians
- Clinicians:
  - Profile: full responsibility for patient care in terms of diagnostics, therapy selection, utilization of tools throughout patient care. Poor IT, software and computational skills
  - Roles in the context of VHT utilization:
    - VHT as a decision support system to support decisions on patient specific therapeutic strategies with focus on optimization. Off-line use of VHT
    - VHT as medical product integrated in established clinical workflows (*e.g.* model-based controller) => on-line use of VHT
    - VHT as alarm system to raise awareness of doctors / nurses on upcoming syndromes in individual patients => on-line use of VHT
- Medical Researcher (can run cohort analysis):
  - Profile: Researcher would like to use VHT for clinical discover.

    - Roles in context of VHT: Access to anonymized data sets and research tools to conduct studies and contribute to advancements in medical knowledge.
- Medical Educator:
    - Profile: Educator aiming to use VHT for education of future professional
    - Roles in context of VHT: Access to VHTs for teaching and training purposes, allowing students and professionals to learn and practice in a simulated environment.

Creator/Model Developer Category

This category of users can upload new models or algorithms, maintain and train existing models or algorithms.

Industrial : In the context of VHT, various types of industrial users who are active in the area of computational medicine / digital twins are expected to use and/or develop digital twins based on the VHT platform. All of them have specific business models and regulatory / IP requirements with direct relation to licensing / reimbursement models, which will be sketched below:

- Solution providers (Software solution)
    - Profile: Companies providing application-specific solutions for patient-specific health care. Development of ready-to use solutions for non-IT experts (*e.g.* clinicians). Often Start-Ups or in-house business units
    - Roles in VHT utilization: need access to data and computational tools
- Consultants
    - Profile: use existing solutions (ready-to use VHT's) to support clients in installation of client-specific solutions, act typically on basis of billable hours + license fees
    - Roles in VHT utilization: 'Power users' of established solutions
- Providers of medical technology:
    - Profile: provide ready to use solutions (mostly hardware) to clinicians /practitioners. VHT has the role as an embedded tool to realize additional value to the established solutions (continuous innovation of product) or enable novel applications significantly exceeding the state of the art in health care (breakthrough innovation)
    - Role in VHT utilization: companies are often owners of huge, but specific patient data sets. Need access to data complementing own data as well as computational environment. Will need to extract application specific solution as dedicated software package enabling integration into hardware solutions.
- Pharma/Biotech
    - Profile: use existing solutions as well as develop new solutions for clinical trial design and real-world evidence analytics
    - Role in VHT utilization: standard clinical trial design / RWE analytics is focussing on cohorts. VHT can play a key role in case of Theranostics concepts as well as individualized therapies, *e.g.* cell therapies (CAR-T). In the latter case dedicated in-house and external experts will develop application specific solutions based on VHT computational platforms. Data may be provided by the companies (based on clinical studies).
- Health Care providers
    - Profile: Health care insurances
    - Role in VHT utilization: Ownership of huge data with very high privacy issues. Focus on VHT development supporting prevention of severe syndromes with specific features for individual customers.

Academic: Academic users are expected to play an important role in further development of computational technologies for next level health care of the future. According to the exponential evolution of the field a definition of the academic users must be rough at the moment.

- Medical/Bio- Informatics
    - Profile: research and development of medical data management and analysis systems
    - Role in VHT utilization: need for access to data, contribution of methods and software components to ecosystem
- Computational Scientist / Systems Medicine
    - Profile: research and development of methods to translate data to knowledge with medical relevance (for clinical use or clinical / biomedical research) by means of models of any type

- Role in VHT utilization: need for access to data, contribution of methods and software components to ecosystem.

Patients/Citizens Category

- Patients/Citizens:
  - Profile: Individuals without background healthcare knowledge
  - Role in VHT utilization: Access to their own VHT, personal medical data, and treatment options, enabling them to better understand their health and make informed decisions about their care.
- Patient Advocate:
  - Profile: Individuals who are empowered with the right to access some individual patient's data.
  - Role in VHT utilization : Access to a specific patient's VHT and medical data with the patient's consent, allowing them to provide support and guidance in healthcare decision-making.
- Citizen Scientist:
  - Profile: Individuals with limited background and ambition to perform citizen/community driven research.
  - Role in VHT utilization : Limited access to anonymized data sets, synthetic data and research tools to contribute to community-driven medical research initiatives.

7.2.2 User incentives and value proposition

Drawing from social sciences approaches and market research techniques, the different user communities are being probed with the aim of identifying types of, and quantify, unmet needs for existing and potential use cases. This will consist of a general survey, accompanied by in-depth interviews with a more limited group. The survey should focus on both the intermediary and end users of the VHT, and on the communities providing assets to the repository. An assessment of the possible financial and non-financial incentives for contributions to the repository can elaborate on different remunerations.

As a first step, representatives of the developer communities were interviewed on the value proposition and incentive mechanisms for developers/researchers to make their resources available on the repository. For the **industrial users**, the propositions change depending on the type of company (pharma, medical devices, software, CRO), their size/maturity (start-up/scale-up/SME & big industry) and the person answering the questions (developers vs executives). In summary, the following (non-exhaustive) list of high-level value propositions was identified

- Use resources (data, models) available in the platform
- Use publicly available portion as sandbox to test new developments
- Use as benchmark
- Facilitate linking companies' own developments with other partners
- Facilitate finding new commercial opportunities
- Benefit from supporting work on technology development
- Benefit from supporting work in establishing ethical, legal and social clarity & certainty

For **academic users**, the aforementioned points also hold true, while additional incentives are specific to the academic case (non-exhaustive)

- Create reward system where uploads and executions can lead to citations
- Use VHT to comply with the local data management system
- Provide private use of the platform for research groups
- Create services that will aid the user such as annotation services, services that will aid science communication etc.

7.3 Evolutionary Ecosystem

Considering the intricate nature of the involved community of practice and their diverse expectations regarding the value propositions of the VHT, the ecosystem will be an evolving and flexible entity. As discussed in Section 4.3.4, to simplify the explanation without oversimplifying its essence, we can delineate three distinct phases that build upon the initial realization of a Distributed Ledger Technology (DLT) system[218] within the VHT infrastructure. This DLT system would enable the permanent tracing of all types of assets exchanged on the platform, while also tracking their origins. Additionally, it would ensure the discoverability, accessibility, semantic interoperability, and reusability of all activated resources. During the initial stages, the DLT infrastructure would exclusively accommodate pre-competitive transactions. Incentives would be based on a quality scoring system, referred to as the "**Honour ledger**". Moving to the second stage, both pre-competitive and competitive transactions would coexist, facilitated by the issuance of digital tokens by the DLT infrastructure. These tokens would possess no direct monetary value but would serve as a framework upon which symbolic prices can emerge through the dynamics of supply and demand for all traded assets, including DLT services. This stage is referred to as the "**Token ledger**". Finally, in the third and final stage, the ecosystem would mature and specialize. While some entities would continue to engage primarily in pre-competitive transactions, an increasing number of participants would shift their focus towards competitive business-to-business exchanges, where prices would be set in Euros rather than tokens. This phase is denoted as the "**Money marketplace**". These phases are discussed in more detail below.

7.3.1 Honour ledger phase

The currency used in this phase is only a "quality" score, based on the reciprocal assignment of quality scoring by all the partners involved. The ecosystem de facto operates as a barter mechanism facilitated by the operation of the DLT infrastructure. Tracing all transactions and quality outcomes, the ledger will be entirely funded with public money.

The main focus in this phase will be on interoperability, *e.g.*, how the participants can barter data and models with the smallest possible effort. The DLT infrastructure should make it very easy to share a dataset or a model in ways compliant with the FAIR principles.

The main motivations for sharing data and models will be because they are forced to do so by the funders. Still, also because it allows them to barter access to their resources with others they can use, and because the use of resources by others will be tracked, possibly with mechanisms that translate them into citation-based reward systems and quality scoring.

The main motivation for reuse will be the simplification of developing and validating new models. This should be particularly true in developing multi-scale and multi-system models, where the work of one on a sub-system can be entirely re-used.

In this phase, the metric of success is how much data and how many models are shared. This translates into another sub-metric: the pain-gain ratio for contributors. The easier it is to share resources, and the more rewarding it is, the better it is.

The governance will require a reverse "T" model: the day-by-day operations will be ensured by a single organisation or a small consortium that is paid to do it and take technical decisions in a fairly autocratic way. Any other decision on the infrastructure is taken with a direct democracy model, where all contributors can participate in the decisional processes with an assembly process.

7.3.2 Token ledger phase

Eventually, the ecosystem will see the flourishing of competitive transactions, aside from the pre-competitive ones. The value of both will be expressed in digital tokens issued by the legal entity governing the DLT infrastructure. The "cashing-in" of tokens, by operators contributing resources to the ledger exchange system and the "paying-out" of tokens, by operators purchasing assets offered by others, will allow the development of an increasing nexus of token prices for all transactions taking place through the DLT, which will also begin to charge a token-fee for its services.

The ledger infrastructure will remain a substantial public resource, but its governance will require a more articulated democracy with representation.

[218] El Ioini N & Pahl C. OTM 2018. Lecture Notes in Computer Science (2018); 11230. https://doi.org/10.1007/978-3-030-02671-4_16

The VHT ecosystem will use its growing token economy to experiment with how it can become progressively self-sustained. In a more advanced stage of development, the DLT infrastructure will possibly also engage in analysing how incentives linked to automated assignment and distribution of value can be determined by ML mechanisms valuing different attributes or even through Shapley-value mathematical methods (inspired by Nobel Laureate Lloyd Shapley) determining the only distribution satisfying a collection of properties within a coalition game.
In its basics, the token-ledger phase will be characterised by token-based exchanges replacing barters, and tokens will be issued to whoever contributes resources to the ledger. However, it will also be possible to purchase tokens in exchange for money: this will apply mainly to external entities not having contributed resources to the ledger but wishing to use the DLT facilities.
Further tokens will be gained anytime shared resources will be used, while everybody will pay with the tokens they have accumulated for being allowed to use somebody else's resources.
The main focus in developing the infrastructure will be the quality of service for its many users. Systems must scale to extensive collections and handle a truly distributed system based on various hardware and software providers.
The main motivations for sharing will remain the same as in the previous phase. But the move from an honour ledger to a tokens ledger will render the incentives for what one can get in return for his/her shared resources much more fine-grained and flexible.
The main motivation for reuse will change by the extent to which the VHT ecosystem will now be influenced by the development strategies of research groups and companies that have purchased tokens and are willing to use the DLT.
This, of course, implies guarantees of persistence for the infrastructure. In this phase, the metric of success is how important the VHT becomes in the development strategies of public and private developers.
Because of the need to ensure long-term sustainability, the VHT infrastructure will have to be run by a legal entity, possibly organised as an NGO (*e.g.*, a foundation) or as a joint undertaking between the EC and the major European industrial players, similar to EuroHPC, IHI, *etc.* This organisation will need to ensure the existence of a public segment of the VHT for the not-for-profit researchers, where most interoperability technicalities standards are tested and standardised. But it will also have to favour the creation of fully commercial segments of the VHT, which are certified for interoperability by the leading legal entity. Beyond that, they are operated entirely in a private way, pursuing sustainable business-to-business models.

7.3.3 Market phase

Eventually, the number of transactions triggered by research groups and commercial companies having monetised their access to the DLT-operated ecosystem will possibly end up changing the nature of the latter, attracting a growing number of subjects out of the ledger and into a fully-fledged money marketplace. Academics will remain the artisans who explore the borders of the VHT territory, while the merchants and entrepreneurs increasingly tend to privatise the ecosystem. Will thus the ecosystem dissolve into a mature industrial sector, or will the DLT infrastructure maintain sufficient resilience and attractiveness because of the advanced qualities of the services it will provide?
In any case, one should expect there to be a public VHT for academic research and early pre-competitive developments, supported by the EC like any other research infrastructure, various VHTs for not-for-profit activities, supported by various charitable mechanisms, and several commercial VHT infrastructures that provide B2B services to an ever-growing industry of in silico medicine.

7.4 Sustainability

7.4.1 Market place services

Moving towards a **sustainable VTH marketplace** requires the provisioning of key services that will ensure the active participation of key stakeholders. One important example of such service lies in the design of mechanisms for **resource valuation**. At a high level, economic incentives play a key role in

any marketplace by ensuring principled, fair valuation of resources, and allocation of value to participants. It is also important, however, to view such mechanisms in a broader context, where "valuation" is not necessarily tied to economics but, for example, may provide the means to ascertain the *quality* of a given resource (data set and/or model). Such quality measures are critical in any vibrant, large-scale ecosystem with potentially numerous available resources addressing the same/similar needs: quality-based rankings can help quickly identify the "best" resources for a given study or task.
The problem of **data valuation** is well studied, and several mechanisms have been proposed. Most existing mechanisms are ad-hoc in nature, including query-based schemes (on-demand or flat pricing), auction-based schemes (using bids to price), and reputation-based schemes (based on user feedback). In contrast, the Shapley Value (SV) mechanism (proposed by Economics Nobel Laureate Lloyd Shapley in 1953) provides a formal framework for data valuation based on viewing Machine Learning (ML) from the data as a coalitional game, with different data sources seen as players in a coalition. In a nutshell, the key idea is to characterize the usefulness of a data source via the ML utility function and to distribute the total value generated by the coalition based on the marginal usefulness of the data sources. SVs can be formally defined through mathematical equations expressing the marginal benefits of a source with respect to improving the target ML utility function and, in fact, give the *only* value distribution mechanism that satisfies some key properties (including task dependence, cumulativeness, fairness, and interdependence of sources). For the practical application of SVs to large-scale marketplaces, various issues need to be addressed including that of computational efficiency, as SV computation is, in general, an exponential operation; still, various methods of efficiently approximating SVs can be exploited (*e.g.*, using sampling).
For the problem of **model valuation**, a large number of model validation metrics have been proposed in the ML literature; these include, for instance, model accuracy (*e.g.*, training or generalization error), speed, memory usage, cross-validation metrics, model complexity (*e.g.*, number of parameters), fairness, explainability, composability, and so on. To provide means for relative model valuation in a VTH marketplace, standard VTH model benchmarks will need to be developed by the community and stakeholders, that can rank such models along these different validation metrics/dimensions. (Similar needs have led to the development of such benchmarks in various other technical communities, including ML, databases, and computer architecture.)
Resource valuation measures can also draw on other important factors, including *provenance*; for instance, a high-quality data set (or, model) can come through the evolution/combination of existing high-quality resources. In fact, this makes another strong argument for the use of DLTs as a means of systematically tracking data and model provenance in the VTH marketplace.
Finally, on a more macroscopic scale, different **market regulation/intervention policies** may be explored to **control the strategic interactions of stakeholders and decision makers** in the VTH marketplace. Such policies could exploit results and theories from technically rich disciplines, such as non-cooperative game theory (e.g., Nash equilibria).

7.4.2 Business models

As a starting point for the development of a proposed business model, we propose to use the Business Model Canvas, a widely used tool in the start-up environment that provides a broad vision of incoming and outgoing cash flows, the value proposition provided by a company, its customers and the relationship with them, the way in which collaborations can be made, and the key activities and resources that can demonstrate the added value of the company/organization involved in VTH. Other tools will be considered to verify the sustainability of a particular business model as for example the SWOT matrix. A focus will be made on their economic vulnerability.

Our initial approach has been to discuss with each partner in charge of a Use Case their respective Business Models, following the Business Model Canvas. Out of our discussion, the following have initially emerged:
1. **Value Proposition**: Virtual twins in healthcare leverage advanced technology, such as artificial intelligence and machine learning, to create digital representations of patients or medical scenarios. These virtual twins offer several value propositions, including personalized treatment planning, predictive analytics, training, maintenance, and education.

2. **Customer Segments**: The target customer segments could include hospitals, medical staff, physicians, medical research institutions, pharmaceutical companies, medical device manufacturers, healthcare providers, and individual patients.

3. **Key Activities**:
- Develop and maintain the virtual twin technology platform, including data collection, analysis, and simulation capabilities. Collecting data in the field can help continuously develop the existing AI models on which the virtual twins are based and can lead to a continuously improved model performance.
- Collaborate with healthcare professionals and researchers to identify specific use cases and tailor virtual twin solutions accordingly.
- Run models through local IT infrastructure or HPC centres.
- Continuously optimize and update the virtual twin models to incorporate new medical knowledge and advances in technology.
- Ensure data security and privacy protocols are in place to protect patient information.
- Ensure maintenance.

4. **Key Resources**:
- Technology infrastructure for data collection, storage, and analysis.
- Expertise in data science, artificial intelligence (machine learning, stochastic processes, predictive algorithms).
- Regulatory compliance and adherence to healthcare industry standards.

5. **Channels**: Channels to reach and engage with the foreseen target customers could include direct sales teams, partnerships with healthcare organizations, online marketing, and participation in industry conferences and events.

6. **Customer Relationships**:
- Provide personalized support and training to healthcare professionals and researchers using the virtual twin technology, including tailoring the virtual twin and underlying model to the specific needs of the customer/user.
- Offer ongoing customer support, addressing technical issues and providing assistance with data interpretation and analysis.
- Foster relationships through regular communication and feedback loops to understand evolving customer needs and enhance the virtual twin solutions accordingly.

7. **Revenue Streams**:
- Licensing or subscription fees (*e.g.* pay-per-use or annual subscription access) for healthcare organizations to access and use the virtual twin platform.
- Consultancy services for customizing virtual twin solutions to specific healthcare use cases.
- Data analysis services
- Integration with electronic health record systems
- Partnership development with pharmaceutical or medical device companies.

8. **Cost Structure**:
- Research and development costs to continuously enhance the virtual twin technology (data quality and volume, models, and algorithms).
- Approvals and certifications (EMEA/FDA, MDR, CE or national mark, ISO standards, HRA)
- Clean data production and acquisition
- Infrastructure costs for data storage, security, and computational resources.
- Medical staff involved in the virtual twin monitoring, repairing, or use
- Sales and marketing expenses to promote the virtual twin solutions.
- Ongoing maintenance and support costs.

9. **Key Partnerships**:
- Collaboration with healthcare institutions, universities, and research organizations to gain access to patient data and validate the virtual twin models.
- Partnership with technology providers for secure data storage, cloud computing, and data analytics.

- Partnership with start-ups with expertise in deployment, integration, and UX/UI design
- Advisors from different fields: healthcare industries, hospitals, and clinics, engineering experts (*e.g.* bioengineering, computational engineering, AI)
- Establishing relationships with pharmaceutical companies and medical device manufacturers to integrate virtual twin technology into their research and development processes.
- Technology Transfer and Innovation Centres: to ensure regulatory compliance and adherence to healthcare industry standards with the help of external Intellectual Property experts

10. **Key Metrics**:
- Size of potential market and quantification of the benefit (socio-economic, ethical, usage and practice, monitoring) that using the virtual twin offers. Specific attention to estimating the number of patients that could be supported and the economic benefit of treating them with the support of the digital twin.
- Number of target customers and user adoption rates.
- Revenue generated from virtual twin platform licenses, subscriptions, and services.
- Customer satisfaction and feedback.
- Accuracy and performance metrics of virtual twin models.
- Return on investment for healthcare organizations utilizing the virtual twin technology.

From these elements, a variety of business models are possible, depending on the maturity level of the specific Digital Twin and on the economic situation:
- Tech transfer or spinoff businesses: starting from university and research lab, going towards ad-hoc created spinoffs, and then to mainstream big Pharma companies
- Specific tech tools: from private deep tech companies specialized in specific fields that provide optimization or support in the use of a particular Digital Twin
- Intermediary businesses: everything that brings assistance or acts as a third party at each step of the VHT process, country and economy-dependant

Some of the more innovative business models that can be foreseen would include:
- Collaborative Ecosystems: Foster partnerships and collaborations with complementary businesses, start-ups, and industry stakeholders. This can help access new markets, share resources, and co-create innovative business models that offer unique value propositions.
- Platform and Marketplace Models: Create a platform or marketplace that connects buyers and sellers, enabling transactions, collaboration, or sharing of resources. This model can unlock new revenue streams, facilitate ecosystem growth, and create network effects.
- Embrace Digital Transformation: Leverage the power of technology and digital solutions to disrupt traditional business models. Explore opportunities to integrate artificial intelligence, machine learning, blockchain, Internet of Things (IoT), and other emerging technologies into your business model.
- Outcome-Based Models: Shift from selling a product or a service to delivering an outcome or value-based solution. This model aligns the success of the virtual twin with the success of the customers by charging based on the results achieved or the value delivered. It encourages innovation and a focus on patient well-being and satisfaction.
- Personalization and Customization: Tailor offerings to meet individual customer preferences and needs, like existing procedures and guidelines. Leverage data, analytics, and technology to deliver personalized experiences, products, or services. This can enhance customer satisfaction, loyalty, and differentiate a business from its competitors.

7.4.3 European Public infrastructure

** In the second half of the EDITH project, different options and modalities of the various EU infrastructure scenarios will be investigated and discussed to suggest relevant options for the VHT initiative and infrastructure.**

7.5 Conclusion

TBD

8 Discussion & Recommendations

The Virtual Human Twin, the digital representations of human physiology and pathology, has considerable potential for medical research and healthcare delivery, enhancing our understanding of human physiology, pathology, and disease aetiology, as well as enabling personalised, patient-centred medicine. In healthcare, DTHs can enable enhanced diagnosis and personalised interventions of higher efficacy and safety across the continuum of prevention, treatment and follow-up. In clinical research, DTHs can also dramatically accelerate development of new medicines and medical devices. Development, validation, and adoption of the VHT will thus advance and promote better healthcare services, improved patient outcomes and efficiency gains for healthcare systems. Adoption of the VHT and building the European VHT ecosystem now will contribute to delivering future innovations required to handle the growing pressures within our healthcare systems.

However, realising the vision for the VHT in health requires addressing several scientific, technical, ethical, legal, and cultural challenges. Delivering the benefits of the VHT for human health requires a robust ecosystem approach enabling researchers, innovators, healthcare practitioners and patients to build on solid foundational research within well-defined frameworks, based on good practices and successful international collaborations. Addressing this vision will initially require a focus on foundational research and the integration of digital and medical technologies towards a shared VHT framework. Respect for citizens' and patients' fundamental rights, and alignment of contributions from stakeholders, will need to be ensured. VHT research and innovation, building upon current achievements, must be oriented towards establishing the VHT as a platform technology, generating evidence and value for healthcare and society as a whole. An ecosystem ensuring incentives for excellence, regulatory certainty and trust will be instrumental in unleashing investment supporting innovation.

In order to realise the vision of the digital twin, the following research and infrastructure challenges need continued and continuous support from European policy makers (non-exhaustive list)

- Excellence in European research and innovation in the development, testing, validation, and verification of advanced VHT technologies, in synergy with existing digital services and capabilities available at European level.
- Advancing the understanding of how VHT solutions, products, and services can be used across the disease continuum, in prevention, treatment and follow-up, in biomedical studies, clinical studies and different healthcare settings, including therapy development, diagnostics, remote care and self-care, as a basis for their development, validation and adoption.
- Identification, development and delivery of high-impact clinical and scientific use cases that stand to benefit from the adoption and use of VHT technologies, products and services, including diagnostics, medical education, training, decision support, therapy development and intervention planning.
- Generation of clinical, experimental and digital evidence for the future development of VHT solutions, methods, and tools and technologies that serve to grow VHT maturity, confirm patient benefit, and increase VHT adoption in healthcare.
- Designing, building, and enhancing the VHT resource repository and simulation platform with expertise and resources in full compliance with applicable laws and regulations in Europe.
- Contribute to the development, testing and implementation of advanced and interoperable IT platform architectures combining novel computational advances, state of the art cybersecurity, cloud services and edge infrastructure supporting the use of advanced VHT models by the citizen, patients, healthcare professionals, researchers and innovators.

- Advancing the availability of and access to high quality, annotated and interoperable digital health data that are standardized in terms of format and semantics, while safeguarding patients' privacy, personal data, health and safety.
- Development of common ground, trust, agreement and certainty on intellectual property management, as a basis from where partner collaborations among stakeholders in developing VHT technologies ensue.
- Identification of opportunities, approaches, standards, tools, and techniques that enhance clarity of the regulatory landscape and support its evolution towards enabling more robust VHT technologies in terms of efficacy, safety, trustworthiness, performance and risk management.

- Ensuring the contributions, feedback, priorities, requirements, views, concerns and interests of citizens, patients, healthcare professionals, and scientists are proactively captured and addressed as part of the development, testing, verification, and validation of VHT technologies.
- Ensuring that VHT technology benefits people of all ages, genders, ethnicities, socioeconomic statuses, and disabilities, fostering equitable and universal access to high-quality healthcare across Europe and worldwide.
- Realising the embedding of the VHT as a European Digital Research Infrastructure providing a continuous source of support and ecosystem development.

The realisation of the development, uptake and immersion of the VHT concept, technologies and infrastructure to advance personalised medicine effectively contribute to a more sustainable and digital future, should be considered as an EU flagship program. We anticipate the program to run over the next ten years (figure 7), with different phases reflecting the development and maturation of the technology, uptake and infrastructure, focusing on realising early and sustained benefits for all stakeholders.

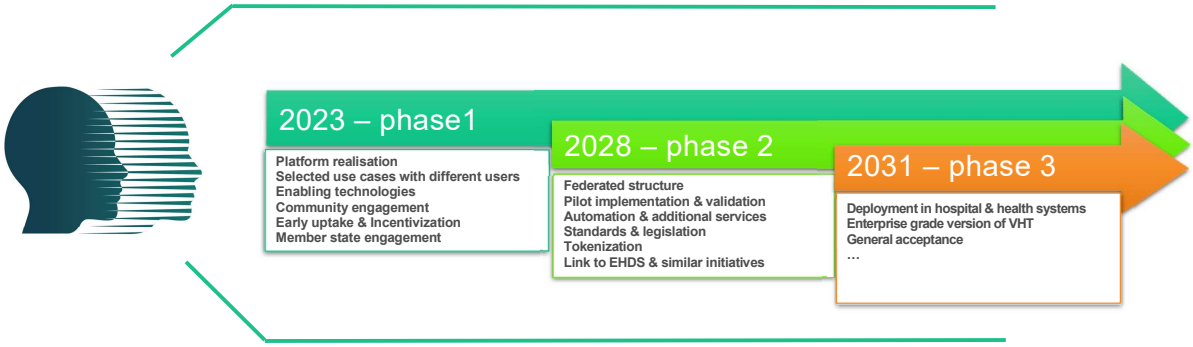

Figure 7: timeline of the VHT roadmap. Tentative starting years of the different phases (reflecting increasing maturity) are provided. Mature elements will pass on from one phase to the next but continuous work will take place for new developments in each phase.

The Virtual Human Twin will facilitate the development and adoption of digital twins in healthcare of any complexity at reasonable costs at reasonable times. As data acquisition and computing power technologies evolve, the scope for digital twins will become broader and the knowledge deeper. This will boost scientific research and technological innovation, creating massive business opportunities in the European Union. The Virtual Human Twin is science, not science fiction; it is the future of medicine.

9 Epilogue: next steps of the EDITH project

As stated before, this is a **preliminary version of the first draft of the VHT roadmap**, reflecting the ongoing discussions within in the consortium and with the external advisory groups- and experts (industry, clinical *etc.*) and the wider community. This is to give all stakeholders the opportunity to ask questions, make remarks and identify potentially missing elements. This input will be combined with the ongoing work of the consortium to create the **first full draft of the VHT roadmap that will be published on the 31st of July 2023**. After publication of this deliverable, it will be disseminated through the established EDITH communication channels (website, social media, newsletter). A form will be available on the project website to collect feedback (www.edith-csa.eu/materials).

After **summer 2023, the public phase** of the project will start. This entails a number of on-site ecosystem meetings in Q4 2023 (likely in Paris) and Q3 2024 (likely in Amsterdam) and other public events as well as public elaboration of the **final version of the VHT roadmap**. Additionally, detailed documentation and guidance will be provided on how to **contribute early prototype demonstrator use cases** and **include resources into the developing repository**. Communication of these initiatives and events will be done through social media, the EDITH website as well as newsletters and other communication channels from all partner institutions. The VHT roadmap public channel on the In Silico World Community of slack will remain open to continue public discussion on specific elements of the roadmap and the infrastructure.

In addition to the further development of the roadmap as a community-driven activity, specific activities are foreseen to **gather further specific input for the roadmap**

- Interact with large (EU) initiatives operating in the healthcare domain, data domain or using/offering technologies that might be relevant for the VHT to identify possible synergies
- Increase interactions with the healthcare providers and hospitals to ensure alignment
- Increase interaction with standardization and regulatory actors/organisations as well as legal actors to facilitate VHT development and use
- Interact with international (non-European) initiatives moving in similar directions.

10 Annexes

10.1 Tables

Table 1: Technical standard defining organisations

Standard defining organisation	Description
ASME[219]	American Society of Mechanical Engineers Defines codes and standards for mechanical engineering. Important subcommittees are: VVUQ SC 40 Computational Modelling of Medical Devices VVUQ SC 60 Guideline for Simulation Software Selection VVUQ SC 70 Machine Learning Applied to Mechanistic & Process Modelling
CEN / CENELEC[220]	European Committee for Standardization The CEN Technical Committee CEN/TC 251 defines standards for health informatics
COMBINE[221]	Computational modelling in Biology Network. Coordinates the development of community standard formats for systems biology modelling.
DICOM[222]	Digital Medicine and Communications in Medicine DICOM is a standard for medical imaging data (ISO 12052:2017).
GA4GH[223]	Genome Alliance for Genomics and Health Standards for collecting, storing, analysing, and sharing of genomic data within research and healthcare
HL7[224]	Health Level 7 (since it is located on the highest level 7, the application layer of the ISO/OSI model communication model) Defined a set of standards for the exchange of electronic clinical and health administrative information
IEC[225]	International Electrotechnical Commission
IEEE[226]	Institute of Electrical and Electronics Engineers
ISO/TC 215[227]	International Standards organisation (ISO) Technical Committee 215 Standardisation in the field of health and medical informatics
ISO/TC 276/WG5[228]	International Standards organisation (ISO) Technical Committee 276 (meta)data standards for the life sciences and for data processing

[219] https://www.asme.org/codes-standards
[220] https://standards.cencenelec.eu
[221] http://co.mbine.org
[222] https://www.dicomstandard.org
[223] https://www.ga4gh.org
[224] https://www.hl7.org
[225] https://www.iec.ch/homepage
[226] https://standards.ieee.org
[227] https://www.iso.org/committee/54960.html
[228] https://www.iso.org/committee/4514241.html

Table 2: Clinical standard defining organisations

Standard defining organisation	**Description**
CDISC[229]	Clinical Data Interchange Standards Consortium Definition of several standards for clinical trials and case report forms (*e.g.*, ADaM, CTR-XML, ODM-XML, OMOP, SDTM)
C-Path[230]	Critical Path Institute
ECRI	Emergency Care Research Institute
ICH[231]	International Council for Harmonisation
IDMP[232]	Identification of Medicinal Products
IHE	Integrating the Healthcare Enterprise
IHTSDO[233]	International Health Terminology Standards Development Organization, the organisation defining the SNOMED-CT[234] (Systematised Nomenclature of Medicine – Clinical Terms) terms
NCI-EVS[235]	National Cancer Institute - Enterprise Vocabulary Service
OHDSI[236]	Observational Health Data Sciences and Informatics
Regenstrief[237]	The Regenstrief institute defines the LOINC and UCUM codes
TransCelerate[238]	TransCelerate Clinical Data Standards Initiative
WHO	World Health Organization

[229] https://www.cdisc.org/standards
[230] https://c-path.org/news-and-events/data-standards/
[231] https://www.ich.org/page/ich-guidelines
[232] https://www.fda.gov/industry/fda-data-standards-advisory-board/identification-medicinal-products-idmp
[233] https://confluence.ihtsdotools.org
[234] https://www.snomed.org
[235] https://evs.nci.nih.gov
[236] https://www.ohdsi.org
[237] https://www.regenstrief.org
[238] https://www.transceleratebiopharmainc.com/initiatives/clinical-data-standards/

Table 3: Standards for medical imaging

<table>
<tr><th>Medical imaging standard</th><th>Description</th></tr>
<tr><td>BIDS[239]</td><td>Brain Imaging Data Structure, a standard of the INCF[240] (International Neuroinformatics Coordinating Facility) for capturing data and metadata of MRI datasets. For different imaging techniques the following extensions[241] to BIDS are available:
<table>
<tr><td>ASL-BIDS</td><td>- for arterial spin labelling</td></tr>
<tr><td>EEG-BIDS</td><td>- for electroencephalographic data</td></tr>
<tr><td>iEEG-BIDS</td><td>- for intracranial EEG data</td></tr>
<tr><td>fNIRS-BIDS</td><td>- for Near InfraRed Spectroscopy</td></tr>
<tr><td>Genetics-BIDS</td><td>- for genetic data associated with human brain imaging</td></tr>
<tr><td>MEG-BIDS</td><td>- for magnetoencephalographic data</td></tr>
<tr><td>Microscopy-BIDS</td><td>- for microscopy imaging data</td></tr>
<tr><td>PET-BIDS</td><td>- for positron emission tomography</td></tr>
<tr><td>qMRI-BIDS</td><td>- for quantitative MRI data</td></tr>
</table></td></tr>
<tr><td>DICOM[242]</td><td>Digital Imaging and Communications in Medicine, a standard for medical imaging-based modelling; stores the data as 2D layers</td></tr>
<tr><td>CIfTI-2[243]</td><td>Connectivity Informatics Technology Initiative (grayordinate surface + volume)</td></tr>
<tr><td>GIfTI[244]</td><td>Geometry (surface) file format</td></tr>
<tr><td>NIfTI-2[245]</td><td>Neuroimaging Informatics Technology Initiative: 2nd Version of a special image format for neuroimaging, where the data are stored in a true 3D volume format</td></tr>
<tr><td>NIfTI-MRS[246]</td><td>extension of NIfTI for magnetic resonance spectroscopy</td></tr>
</table>

[239] https://bids.neuroimaging.io
[240] https://www.incf.org
[241] https://bids.neuroimaging.io/get_involved.html#extending-the-bids-specification
[242] https://www.dicomstandard.org
[243] https://www.nitrc.org/forum/attachment.php?attachid=333&group_id=454&forum_id=1955
[244] https://www.nitrc.org/projects/gifti/
[245] https://nifti.nimh.nih.gov/nifti-2
[246] https://wtclarke.github.io/mrs_nifti_standard/

Table 4: Standard formats for electro- and neurophysiology, biosignal and vital sign data

Standard	Description
ECG-XML	An XML-based format for electrocardiogram data
EDF / EDF+[247]	European Data Format, a format for exchange and storage of multichannel biological and physical signals, *e.g.*, polysomnography
GDF[248]	General Data Format; for biomedical signals, like *e.g.*, EEG and ECG data
HL7-aECG[249]	HL7-annotated Electrocardiogram; an annotated XML-based format for electrocardiogram data
ISO/IEEE 11073[250]	A standard for device interoperability of point-of-care medical device communication
MEF[251]	Multiscale Electrophysiology Format for EEG data
MFER[252]	Medical waveform Format Encoding Rules, a file format for encoding medical waveforms from ECG, ECoG and EEG data
NIX[253]	Neuroscience Information eXchange for neurophysiology data
NSDF[254]	Neuroscience Simulation Data Format based on HDF5
NWB:N 2.0[255]	Neurodata Without Borders:Neurophysiology for neurophysiology data
OpenEP[256]	A cross-platform electroanatomic mapping data format
SCP-ECG[257]	Standard Communications Protocol for ECG data
SignalML[258]	An XML-based meta-format for biomedical time series data
SONATA[259]	Scalable Open Network Architecture TemplAte; for large-scale modelling of neuronal brain network models and simulation output
VSIR[260]	Vital Signs Information Representation (CEN 13734)
WFDB[261]	WaveForm DataBase format of the PhysioNet (physio.net) repository; a combination of MIT format and EDF / EDF+

[247] https://www.edfplus.info
[248] http://justsolve.archiveteam.org/wiki/General_Data_Format_for_Biosignals
[249] https://en.wikipedia.org/wiki/HL7_aECG
[250] https://www.iso.org/standard/77338.html
[251] https://main.ieeg.org/?q=node/28
[252] http://www.mfer.org/en/index.htm
[253] http://g-node.github.io/nix/
[254] https://github.com/nsdf/nsdf
[255] https://www.nwb.org/2019/02/26/nwbn-2-0-final-released/
[256] https://openep.io
[257] https://github.com/topics/scp-ecg
[258] https://braintech.pl/software/svarog/signalml/?lang=en
[259] https://docs.sonata-project.org/en/master/
[260] https://standards.iteh.ai/catalog/standards/cen/8f621d17-ffcc-4885-bfb2-ff72df02e7f1/env-13734-2000
[261] https://wfdb.readthedocs.io/en/latest/wfdb.html

Table 5: Standards for genetic sequence variants

Sequence variant format	**Description**
VCF[262]	Variant Call Format
BCF	Binary Call Format, a binary version of VCF
MAF[263]	Mutation Annotation Format
GVF	Genome Variation Format, an extension of GFF3
GVCF[264]	Genomic Variant Call Format
SPDI	Sequence, Position, Deletion, Insertion
GA4GH-VR[265]	Genome Alliance for Genomics and Health - Variation Representation
HGVS[266]	Human Genome Variation Society, following the HGVS sequence variant nomenclature

[262] https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format
[263] https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/
[264] https://gatk.broadinstitute.org/hc/en-us/articles/360035531812-GVCF-Genomic-Variant-Call-Format
[265] https://vrs.ga4gh.org/en/latest/
[266] https://varnomen.hgvs.org/bg-material/simple/

Table 6: Standards for models

Modelling standard	Description
BioPAX[267]	Biological Pathways eXchange for exchange and visualisation of biological pathway data
CellML[268]	XML-based description and exchange format for cellular models
FieldML[269]	Human Physiome Field Markup Language, an XML-based format using mathematical field descriptions of cells, tissues and organs; can be used to represent finite element models
MDL[270]	Model Description Language for pharmacometric models
MoBi[271]	For multiscale physiological modelling and simulation
MorpheusML[272]	Format for agent-based multicellular models
MultiCellDS[273]	MultiCellular Data Standard for centre-based models (CBMs) [Montagud 2021]
MultiCellML[274]	Standard for agent-based multiscale and multicellular spatial models
NeuroML v2[275]	Neuroscience eXtensible Markup Language; an XML-based description and exchange format for models in neuroscience with its four parts: Biophysics, ChannelML, MorphML, and NetworkML
OpenBEL[276]	Biological Expression Language, a triple-based (subject-predicate-object) modelling language for representing biological knowledge by causal, correlative, and associative relationships
PharmML	Pharmacometrics Markup Language, an exchange format for pharmacokinetic and pharmacodynamic models
PK-Sim[277]	Standard model for whole-body physiologically based pharmacokinetic modelling
SBML[278]	Systems Biology Markup Language, an XML-based description and exchange format for differential-equation models of biological processes. SBML Level 3 is a modular format with a core and packages for extending that core functionality:

SBML-arrays	- for vectorized, *e.g.*, grid-based models
SBML-comb	- for multiscale and modular, *e.g.*, tissue and whole-body models
SBML-distrib	- for distributions, *e.g.*, systems pharmacology and population models
SBML-dyn	- for dynamical models
SBML-fbc	- for constraint-based flux-balance (steady state), *e.g.*, genome-scale models
SBML-multi	- for rule-based, *e.g.*, multistate molecules and multicomponent complexed
SBML-layout	- for visualisation
SBML-groups	- for grouping and organisation
SBML-qual	- for qualitative (*i.e.*, Boolean) models
SBML-render	- for visualisation
SBML-spatial	- for spatial, *e.g.*, reaction-diffusion models

[267] http://www.biopax.org
[268] https://www.cellml.org
[269] https://physiomeproject.org/software/fieldml
[270] http://mdl.community
[271] https://github.com/Open-Systems-Pharmacology/MoBi
[272] https://morpheus.gitlab.io/tag/morpheusml/
[273] http://multicellds.org
[274] https://multicellml.org/wiki/doku.php
[275] https://neuroml.org
[276] https://github.com/OpenBEL
[277] https://github.com/Open-Systems-Pharmacology/PK-Sim
[278] https://sbml.org

SBOL[279]	Synthetic Biology Open Language, describing the exchange of synthetic biological genetic parts, devices, modules, and systems.
VCML[280]	Virtual Cell Markup Language for rule-based modelling

[279] https://sbolstandard.org
[280] https://vcell.org/webstart/VCell_Tutorials/VCell_Help/topics/ch_1/Introduction/Export.html

Table 7: Standards for model simulations and documentation of results

Standard	Description
GSP[281]	Good Simulation Practice, a quality standard for in silico simulations, developed by the Avicenna Alliance
NuML[282]	Numerical Markup Language, an XML-based format for exchanging numerical data
OMEX	Open modelling EXchange for exchange of modelling and simulation data. An OMEX file is a .zip container containing a manifest file, an optional metadata file and the data files.
SBRML[283]	Systems Biology Results Markup Language for encoding results of SBML simulations
SED-ML[284]	Simulation Experiment Description, a XML-based exchange format for encoding of simulation setups following the MIASE guidelines

[281] https://insilico.world/sito/wp-content/uploads/2023/05/Position-Paper-GSP-R6.pdf
[282] https://github.com/NuML/NuML
[283] https://sbrml.sourceforge.net/SBRML/Welcome.html
[284] https://sed-ml.org

Table 8: Terminologies and ontologies for the description and annotation of data, models, and their components

Ontology	Description
BCTEO	Bone/Cartilage Tissue Engineering Ontology
BRICKS[285]	Describes recurring concept in SBGN models
CBO[286]	The Cell Behavior Ontology is designed to describe multi-cell computational models
CheBI[287]	Chemical Entities of Biological Interest
CMDO	Clinical MetaData Ontology
CNO[288]	Computational Neuroscience Ontology
CO	Cell Ontology
COMODI[289]	COmputational MOdels Differ, an ontology to characterise differences in versions of computational biology Models
CVDO	Cardiovascular Disease Ontology
DEB[290]	Devices, Experimental Scaffolds, and Biomaterials Ontology
DINTO[291]	Drug-drug Interaction Ontology
DO[292]	Disease ontology
DTO[293]	Drug target ontology
DUO[294]	Data Use Ontology
ECTO[295]	Environmental Conditions, Treatments, and exposures Ontology
EDAM[296]	EMBRACE Data and Methods ontology
FMA[297]	Foundational Model of Anatomy for localising anatomical structures at a specific spatial location
HPO[298]	Human Phenotype Ontology
HUPSON	Human Physiology Simulation Ontology containing concepts for simulations, models, and algorithms
InChI[299]	IUPAC International Chemical Identifier
KISAO[300]	Kinetic Simulation Algorithm Ontology for describing simulation algorithms
LiCO	Liver case ontology
MAMO	Mathematical Modelling Ontology
MAxO[301]	Medical Action Ontology
MONDO[302]	A semi-automatically constructed ontology defined by the Monarch initiative that merges multiple disease ontologies to yield a coherent merged ontology.
ND[303]	Neural Disease Ontology
NPO[304]	NanoParticle Ontology
NPU	Nomenclature for Properties and Units terminology

[285] https://brickschema.org/ontology/
[286] https://cbo.biocomplexity.indiana.edu
[287] https://www.ebi.ac.uk/chebi
[288] https://github.com/INCF/Computational-Neurosciences-Ontology--C.N.O.-
[289] http://comodi.sems.uni-rostock.de
[290] https://github.com/ProjectDebbie/Ontology_DEB
[291] https://github.com/labda/DINTO
[292] https://disease-ontology.org
[293] http://drugtargetontology.org
[294] https://github.com/EBISPOT/DUO
[295] http://obofoundry.org/ontology/ecto.html
[296] https://edamontology.org/page
[297] http://sig.biostr.washington.edu/projects/fm/AboutFM.html
[298] https://hpo.jax.org/app/
[299] https://iupac.org/who-we-are/divisions/division-details/inchi/
[300] https://github.com/SED-ML/KiSAO/
[301] https://github.com/monarch-initiative/MAxO
[302] https://mondo.monarchinitiative.org
[303] https://github.com/addiehl/neurological-disease-ontology
[304] http://www.nano-ontology.org

OBCS[305]	Ontology of Biological and Clinical Statistics
OBI[306]	Ontology for Biomedical Investigations
OGMS[307]	Ontology for General Medical Science
OPD	Ontology of Physics for Biology containing physics concepts to describe the dynamics of biological systems
OPMI[308]	Ontology of Precision Medicine and Investigation
PAV[309]	Ontology for provenance, authoring and versioning
PK ontology	An ontology for pharmacokinetic models
PreMedOnto	Precision Medicine Ontology
ROO[310]	Radiation Oncology Ontology
SBO[311]	Systems Biology Ontology with terms useful for describing computational modelling
SBOL-VO	Synthetic Biology Open Language Visual Ontology
SIO[312]	Semantic science Integrated Ontology
STATO[313]	Ontology of Statistical methods
TEDDY	Terminology for the Description of Dynamics. An Ontology for the description of control elements and dynamics in systems biology and synthetic biology
UCUM[314]	Unified Code for Units of Measure, defined by the Regenstrief institute

[305] https://github.com/obcs/obcs
[306] https://obi-ontology.org
[307] https://github.com/OGMS
[308] https://github.com/OPMI/opmi
[309] https://pav-ontology.github.io/pav/
[310] https://www.cancerdata.org/roo-information
[311] https://github.com/EBI-BioModels/SBO
[312] https://github.com/MaastrichtU-IDS/semanticscience
[313] https://stato-ontology.org
[314] https://ucum.nlm.nih.gov

Table 9: Clinical languages, terminologies, and code systems

Language, terminology / code system	Description
Arden syntax[315]	a HL7 standard for the representation of medical knowledge, *e.g.*, for use by CDSSs
ATC[316]	Anatomical Therapeutic Chemical code, a classification system for the active substances of biomedical drugs
CQL[317]	Clinical Quality Language, a HL 7 standard for the expression of clinical knowledge used for CDS and electronic Clinical Quality Measurement (eCQM); can also be used for querying complementing FHIR search
EMDN[318]	European Medical Device Nomenclature
GMDN[319]	Global Medical Device Nomenclature
ICD-11[320]	International Classification of Diseases
ICF[321]	International Classification of Functioning, Disability and Health
ICHI[322]	International Classification of Health Interventions
LOINC[323]	Logical Observation Identifiers Names and Codes for reporting laboratory test results
MedDRA[324]	Medical Dictionary for Regulatory Activities
NCIt[325]	the National Cancer Institute thesaurus
ORDO[326]	Orphanet Rare Disease Ontology
ORPHAcode[327]	Encoding of rare diseases and orphan drugs
RxNorm[328]	Normalised Names for clinical drugs
SNOMED-CT[329]	Systematized Nomenclature of Medicine – Clinical Terms
TNM[330]	Tumour Node Metastasis, a classification system for malignant tumours
UMDNS[331]	Universal Medical Device Nomenclature System of ECRI institute

[315] https://www.hl7.org/implement/standards/product_brief.cfm?product_id=2
[316] https://www.ema.europa.eu/en/glossary/atc-code
[317] https://cql.hl7.org
[318] https://webgate.ec.europa.eu/dyna2/emdn/
[319] https://www.gmdnagency.org
[320] https://icd.org
[321] https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health
[322] https://www.who.int/standards/classifications/international-classification-of-health-interventions
[323] https://loinc.org
[324] https://www.meddra.org/
[325] https://ncithesaurus.nci.nih.gov/ncitbrowser/
[326] https://www.orpha.net/consor/cgi-bin/index.php?lng=EN
[327] https://www.orpha.net/consor/cgi-bin/Disease.php?lng=EN
[328] https://www.nlm.nih.gov/research/umls/rxnorm/overview.html
[329] https://www.snomed.org
[330] https://www.uicc.org/resources/tnm
[331] https://www.ecri.org/solutions/umdns