ISW

In Silico World: Lowering barriers to ubiquitous adoption of In Silico Trials

Grant Agreement No. 101016503

D3.1 **Report on initial validation data collections (D6)**

Deliverable information	
WP number and title	WP3 Validation collections
Lead beneficiary	TU/e
Dissemination level	Public
Due date	30/9/2021
Actual date of delivery	09/11/2021
Author	TU/e
Contributors	TU/e

The following document reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016503

Quality assurance

To ensure the quality and correctness of this deliverable, we implied an internal review and validation process. The deliverable was drafted by the work package leader (TUE). All partners contributed to and reviewed the overall draft. Finally, the semi-final version was submitted to the project coordinator, for a final review and validation.

Version	Date	Status	Author	
V0	29/09/2021	Draft	TU/e	Creation of the document
V1	01/10/2021	Draft	TU/e	First local iteration
V2	24/10/2021	Draft	TU/e	Feedback collected from partners
V3		Final version		

Table of contents

1 Introduction

General aim

This deliverable (report) focuses on the collection, curation, and publication of at least 7 data collections designed to provide validation to *In Silico* Trials solutions. The final goal is to provide reusable resources for validation that others can use for setting up their own *in silico* clinical trial and to provide validation collections for a set of specific applications.
This implies to design and implement dedicated **tools to**:

- manage data collections,
- create synthetic data collections,
- define data uncertainty,
- perform sensitivity analysis,
- quantify uncertainty.

These tools will be developed during the ISW project in close collaboration with the partners that own/develop the models.
Eventually, these tools will be applied in a number of pilot showcases and to fully-featured solutions developed in WP2. The initial strategy will be discussed in this report.

The process will start with the curation and preliminary publication of some validation datasets already available; from this early work, it will be possible to define the **requirements** for validation collections and to develop an appropriate certification process. Depending on the quality and completeness of the original data collections provided, the result will guide the development of the final collections, which will include also new datasets specifically curated for the purpose. Lastly, the involved partners will demonstrate the usefulness of these collections in the validation process.

Specific aim

Initially, a data collection to validate coronary artery disease models (**FFRValid**) will be curated and published. This collection will be analysed concerning its intended use and published in a findable, accessible, interoperable, and reusable (FAIR) format. The focus will be on the curation aspects and **data certification**, as described in detail more in section 2 below. In addition to the use of the tools mentioned above, this also includes considering regulatory accreditation (in collaboration with WP4: Technical Standards and Regulatory Aspects), storage facilities (in collaboration with WP6: Scalability and Efficient Computing)), and accessibility models (in collaboration with WP8: Exploitation, ecosystem enlargement, and technology transfer).

In a later stage (tasks 3.2 (M10-M18), with the support of the other partners, a set of requirements and specific guidelines to assist curating and publishing validation collections will be composed. The analysis will identify and separate general requirements from case-specific requirements. In close collaboration with WP4 and WP9, this requirements analysis will also consider the ethical, societal, and regulatory requirements.

2 Methods

Curation

The curation of the data will be based on the structure that is defined as depicted in Figure 1. Based on the legal, ethical, and societal constraints defined in WP4 and WP8, data will be stored using the facilities defined in WP6. This can either be the original data or a synthetic data collection that fully represents the original data. This opens the possibility to keep the original data at the site where they were acquired and under the responsibility of the owners of the data. For the original data, the strategy defined in the data management plan of WP9 (Legal and ethical framework) is followed. Note that a first version of this strategy is described in Deliverable D9.5.

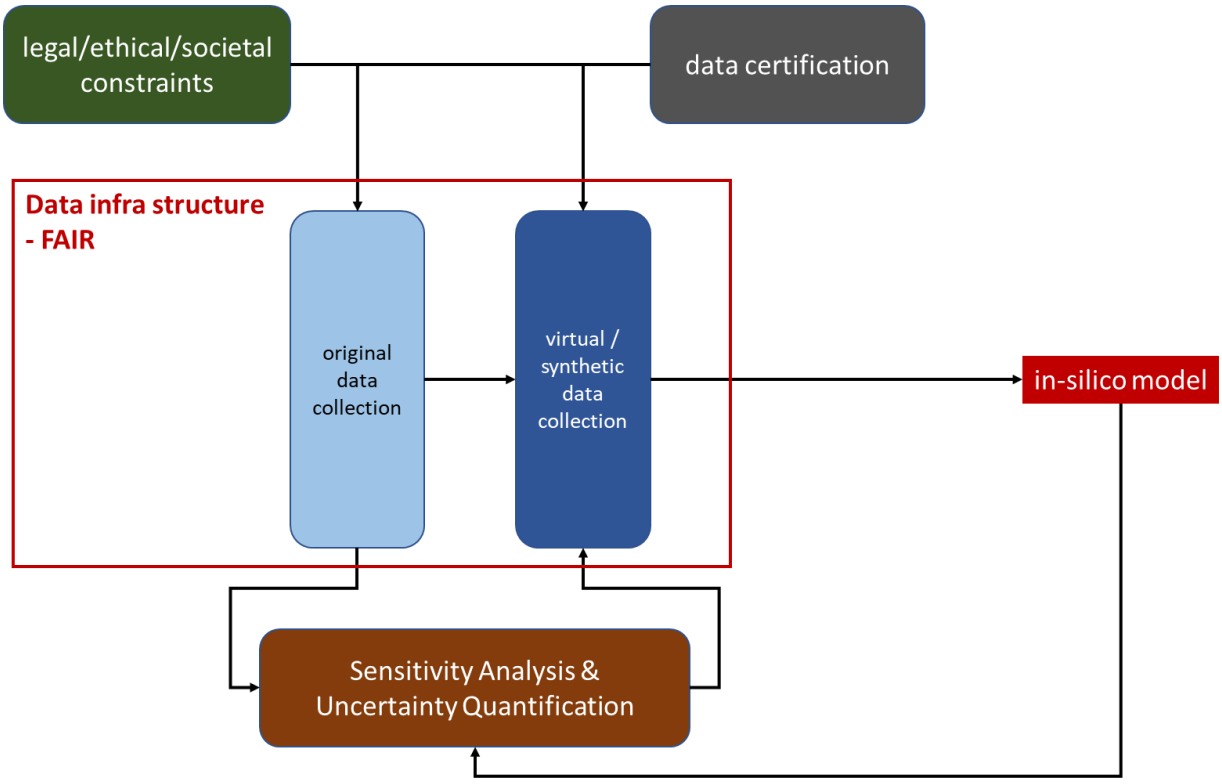

Figure 1: General strategy regarding usage of validation data collections (all activities mentioned will be a collaboration between the different WPs.

Data certification will be described in more detail in the next section of this report. Sensitivity analysis and uncertainty quantification (SA & UQ) will be part of future deliverables and require input from WP2 (Development of Advanced Solutions). We will use **FFRValid** (see description below) as a first pilot to validate this general strategy. First versions of models to predict the outcome of clinical intervention are already available for this application, so initial versions of the complete chain of tools can already be implemented in a relatively early stage of the project.

FFRValid: *in collaboration with the Catharina hospital in Eindhoven the angio-based data that has been gathered in the FAME 1 and FAME 3 clinical trials to validate FFR guided versus angio-based PCI and FFR guided PCI versus CABG respectively. The data collections consist of 1000 and 1500 X-ray angiograms, partly mono-plane and partly biplane combined with measured local FFR at specific annotated branches of the coronary tree. Automated segmentation will be developed and carried out to represent these data in terms of centerlines and local diameter. Patient phenotype and all available and relevant patient record data will be collected in a predefined irreversibly anonymized database structure. The data collection will also be used to develop procedures for data assimilation, generation of virtual patient cohorts, sensitivity analysis and uncertainty quantification, hybrid modelling combining data-driven models with mechanistic models, and meta-modeling.*

Certification

The following challenges in this, and other, data collection can be distinguished:

- Not all clinical parameters are always measured for every patient.
- Clinical measurements are hampered by relatively large uncertainties.
- Most of the data is sparse, i.e., limited data availability for some clinical conditions.

To make the data collections suitable for In Silico Trials (IST) we propose a two-stage approach. In the first stage (see Figure 2), synthetic data will be generated by using three different intermediate steps: Firstly, data imputation is used to account for the missing data within the acquired data set and thus complete it [1]. Secondly, Generative Adversarial Networks (GANs) or similar techniques will be used for the generation of synthetic datasets [2], considering all variabilities so that they can be used for future research as open-source data (see stage 2). Finally, the sparse datasets are augmented to properly account for outliers and special cases within the total data set [3]. The synthetic data could be used for virtual cohort generation using sensitivity analysis and uncertainty quantification [4,5].

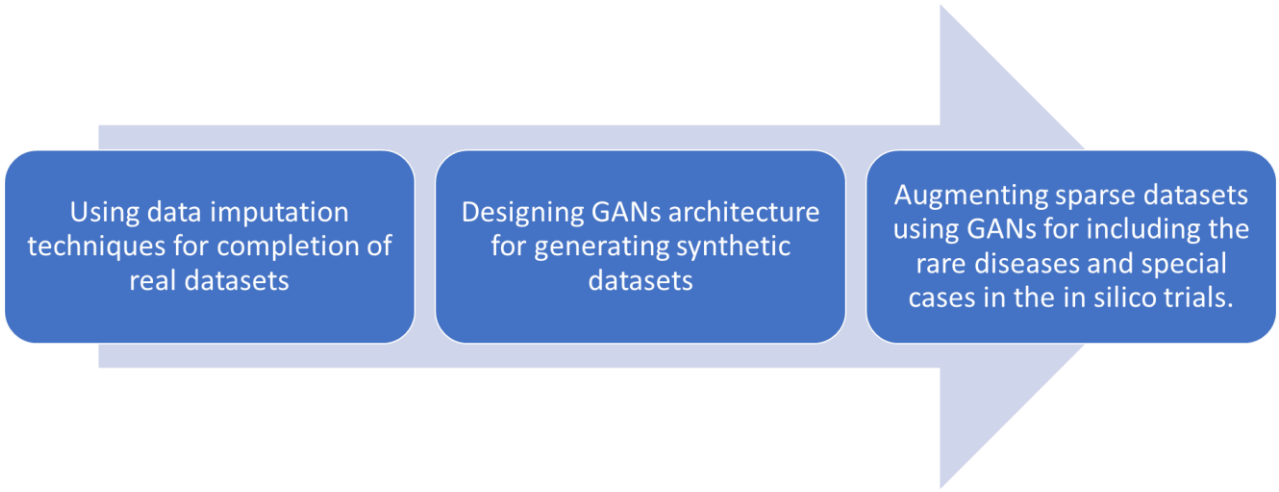

Figure 2: Stage 1: Generation of a synthetic dataset based on GANs

Figure 3: Stage 2: Generation of an operational data set for validation and ISCTs

In the second stage (see Figure 3), an operational dataset will be generated that is fully suitable for ISTs but also can be made available for public use. Again, this is achieved via three intermediate steps: Firstly, the focus should be on acquiring legal documents as the data might contain confidential information. This is obsolete, however, for datasets comprised of only synthetically generated data. Sensitivity analysis and uncertainty quantification will be executed for model-specific parameter prioritization [6]. Secondly, in close collaboration with WP4 (Technical Standards and Regulatory Aspects), the synthetic datasets should be validated to see if they contain all the anatomical and functional details and features present in the real dataset so that they can be used for *in silico* trials instead of going through the tedious process of collecting huge patient datasets. Lastly, in close collaboration with WP6 (Scalability and Efficient Computing) and WP8 (Exploitation, ecosystem enlargement, and technology transfer), the datasets will be made accessible to the public.

Detailed Workplan

For each of the data collections, starting with **FFRValid**, we will start with collecting a single sample data set $S_0(\underline{X})$ (dataset Zero). Dataset Zero must have the structure of a complete dataset of the collection and consists of a list of M entries $\underline{X} = \{\underline{x}^i\}_{i=1}^{M}$. The entries $\underline{x}^i$ can be of any type ranging from a complete image (or series of images, e.g., angiograms) to a single value (e.g. age of the patient). For each of the data collection that will be considered in the ISW project, the completeness of the dataset has to be defined. This will depend on the intended use.

After analysis of the sample regarding its content $\underline{X}$, we will generate two surrogate data collections $\underline{C}_0^g(\underline{X})$ and $\underline{C}_0^p(\underline{X})$, a generic surrogate data collection and a patient-specific surrogate data collection respectively. Generic surrogate collection will be a generalized dataset which can be utilized for any relevant research whereas patient-specific dataset depends on the intended context of use of a model that relies on that data. Each collection consists of a number of data sets, N_g and N_p respectively. The generic surrogate collection will be a set in which the content $\underline{X}$ is varied with a population variance derived from literature. The patient-specific surrogate collection will be based on uncertainty ranges in the entries $\underline{x}^i$ of content $\underline{X}$ due to "measurement" errors or other uncertainties, e.g. uncertainties due to the physiological envelope at which the data is acquired. The surrogate data collections can be manipulated by inducing missing data (create incompleteness) enabling evaluation of data imputation techniques.

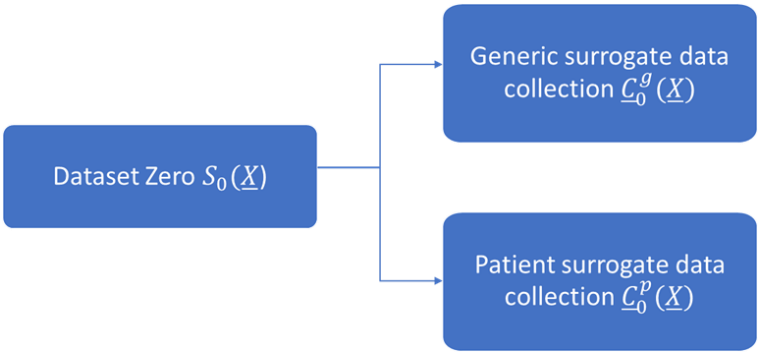

Figure 4: Creation of surrogate data collections

Rare diseases (which may be rare in one demography but are majorly prevalent in other demography) can also be introduced to make a complete synthetic dataset, thereby making the surrogate data collection suitable for universal use. Another advantage of starting with the surrogate data collections is that the tool development can start before all datasets are fully available, even if not all data is available in the correct format and at an accessible location. In a later stage, the surrogate data collections will be "infected/enriched" with real data to obtain representative synthetic collections $C_s^g(\underline{X})$ and $C_s^p(\underline{X})$, consisting of datasets $S_p^{n_p}(\underline{X})$ and $S_g^{n_g}(\underline{X})$ with n_p and n_g representing the number of datasets within the collections respectively. Note that we talk about a patient data collection where the collection refers to realizations of variated input of patient dataset $S_p(\underline{X})$.

The next step is to simulate the specific clinical trial proposed using the models developed in WP2 (Development of Advanced Solutions). In the chain of tools to be developed, we either use the original model or a meta-model. In both cases the model produces output $\underline{Y}$ as a function of input $S(\underline{X})$, the model parameters $\underline{\alpha}$, and the clinical trial optimization/selection criteria $\underline{\beta}$. The outcome $\underline{Y}(\underline{X}, \underline{\alpha}, \underline{\beta})$ needs to be translated to clinical metrics $\underline{\gamma} = \underline{Z}(\underline{Y})$, which are used as endpoints of the clinical trial. Uncertainty quantification is computed based on the sensitivity of $\underline{\gamma}$, which is dependent on $\underline{Y}$. Uncertainty in the model output $\underline{Y}$ is dependent on the variance of $\underline{X}$, $\underline{\alpha}$, and $\underline{\beta}$. This quantification will help in prioritizing the parameters which will lead to more accurate final outcome of the clinical trial.

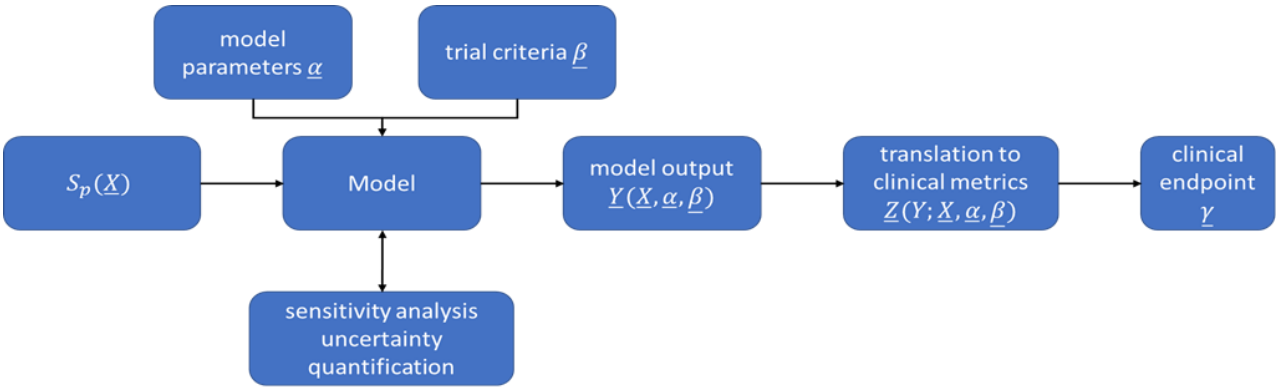

Figure 5: Flowchart to illustrate the in silico clinical trial toolchain adapted from the work of Bodner and Kaul [7]

3 Conclusion

This deliverable aims to provide a strategy for the collection, curation, and publication of several data sets designed to provide validation to *In Silico* Trials solutions. Clear general steps have been defined in order to create suitable synthetic datasets to perform *in silico* clinical trials with regard to data management and data certification. For 5 out of 7 data collections a first phase data management plan has been established and regarding the pilot dataset FFRValid first options for automatic segmentation are now being developed. In addition, the model that will be used to simulate the clinical trial has been re-implemented (from Matlab to Python) in order to make it suitable for use in the ISW project.

References

[1] Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med. 2020 Nov 9;3(1):147. doi: 10.1038/s41746-020-00353-9. PMID: 33299100; PMCID: PMC7653933.

[2] Ghorbani, A., Natarajan, V., Coz, D. & Liu, Y.. (2020). DermGAN: Synthetic Generation of Clinical Skin Images with Pathology. Proceedings of the Machine Learning for Health NeurIPS Workshop, in Proceedings of Machine Learning Research 116:155-170.

[3] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J Big Data* **6,** 60 (2019). https://doi.org/10.1186/s40537-019-0197-0.

[4] Eck VG, Donders WP, Sturdy J, Feinberg J, Delhaas T, Hellevik LR, Huberts W. A guide to uncertainty quantification and sensitivity analysis for cardiovascular applications. Int J Numer Method Biomed Eng. 2016 Aug;32(8). doi: 10.1002/cnm.2755. Epub 2015 Nov 26. PMID: 26475178.

[5] Gabriel D. Maher, Casey M. Fleeter, Daniele E. Schiavazzi, Alison L. Marsden, Geometric uncertainty in patient-specific cardiovascular modeling with convolutional dropout networks, Computer Methods in Applied Mechanics and Engineering, Volume 386, 2021,114038, ISSN 0045-7825, https://doi.org/10.1016/j.cma.2021.114038.

[6] Donders WP, Huberts W, van de Vosse FN, Delhaas T. Personalization of models with many model parameters: an efficient sensitivity analysis approach. Int J Numer Method Biomed Eng. 2015 Oct;31(10). doi: 10.1002/cnm.2727. Epub 2015 Jun 15. PMID: 26017545.

[7] Bodner J, Kaul V. A framework for In Silico Clinical Trials for medical devices using concepts from model verification, validation and uncertainty quantification (VVUQ). *Proceedings of the ASME 2021 Verification and Validation Symposium. AMSE 2021 Verification and Validation Symposium*. May 19-20, 2021.