DELIVERABLE

Project Acronym: **SMASH-HCM**

Grant Agreement number: **101137115**

Project Title: **Stratification, Management, and Guidance of Hypertrophic Cardiomyopathy Patients using Hybrid Digital Twin Solutions**

D7.6 – SMASH-HCM solution as research tool requirements

Revision: 1.0

Authors and Contributors	Malte von Tottleben (EMP), Shahan Tariq (EMP), Marie Reuter (EMP)		
Responsible Author	Malte von Tottleben	Email	Malte.vonTottleben@empirica.com
	Beneficiary	EMP	

Project co-funded by the European Commission within HORIZON-HLTH-2023-TOOL-05-03		
Dissemination Level		
PU	Public, fully open	X
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

Revision	Date	Author	Organisation	Description
0.1	03.07.2025	Malte von Tottleben	EMP	Initial draft created
0.2	29.07.2025	Malte von Tottleben, Marie Reuter	EMP	Added system requirements section
0.3	02.09.2025	Malte von Tottleben, Marie Reuter, Shahan Tariq	EMP	Integrated feedback from consortium partners, SMASH-HCM as a research tool, Introduction and Conclusion
0.4	25.09.2025	Malte von Tottleben	EMP	Pre-final version with editorial revisions
0.9	25.09.2025	Mark van Gils	TAU	Pre-final version review and suggestions
1.0	30.09.2025		EMP, TAU	Final version submitted to EC

Date of delivery	Contractual:	30.09.2025	Actual:	30.09.2025
Status	final ☒ /draft ☐			

Abstract (for dissemination)	This deliverable defines requirements for the SMASH-HCM research toolbox, a modular collection of computational models, algorithms, and APIs supporting hypertrophic cardiomyopathy studies. It specifies functional, technical, ethical, and sustainability aspects, emphasising interoperability, reproducibility, and open science principles. The toolbox fosters translational research, algorithm development, and integration with European digital health infrastructures.
Keywords	Research toolbox, Hypertrophic Cardiomyopathy, Digital Twin, Virtual Human Twin, Open Science, Reproducibility, Interoperability

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Content

Executive Summary

This deliverable D7.6 specifies the requirements for the SMASH-HCM solution in its function as a research toolbox. While earlier deliverables (D7.1 and D7.2) concentrated on clinical and patient-oriented features, the present document changes perspective and addresses researchers, modellers, and academic developers. It considers the scientific need for an open and modular research toolbox that can extend the value of the SMASH-HCM research results beyond the regulated clinical product.

The SMASH-HCM research toolbox is defined as a collection of computational models, algorithms, and APIs, developed and deployed in a containerised and well-documented manner. It is not a monolithic software but a flexible toolbox. Researchers may adapt, combine, and extend components for their own workflows. The toolbox is expected to serve use cases such as in-silico trials, algorithm development, multi-scale simulation, educational training, and industrial innovation. This approach aligns with the Horizon Europe mission to foster reusable digital twin solutions.

The deliverable further details the system requirements for such a toolbox, covering functional, technical, performance, and legal aspects. Examples include integration with EHR systems through HL7 FHIR or OMOP CDM standards, ingestion of research datasets, provision of reproducible analysis environments, privacy-preserving methods such as federated learning, and compliance with GDPR and ethical standards. The research toolbox is expected to prepare for future interoperability with external infrastructures (e.g., EDITH, Virtual Human Twins initiative) and to contribute to open science through accessible repositories, versioning, and clear documentation where useful.

The requirements presented here reflect both current needs and future aspirations. Some features are foreseen within the SMASH-HCM project itself, while others are recognised as desirable for later development. By documenting them in this form, D7.6 provides a blueprint for sustainable continuation of SMASH-HCM as a research toolbox, guiding both immediate implementation and long-term integration into the European digital health research landscape.

1 Introduction

This deliverable outlines functional requirements for the SMASH-HCM solution as a research tool, addressing its non-technical specification in a potential future continued research setting. It complements earlier deliverables focused on clinical and patient functional requirements (D7.1, D7.2), by shifting the perspective to a different category of end-users: researchers, modellers, data scientists, and academic developers. This deliverable defines the functional requirements and potential user expectations for a version of SMASH-HCM intended not for clinical decision-making, but for exploratory, experimental, and translational research. The intended users include those who wish to adapt, extend, or build upon the digital twin models and analytics pipeline created during the project.

Several motivations underscore the importance of this work:

- Scientific and funding relevance: SMASH-HCM was funded under the HORIZON-HLTH-2023-TOOL-05-03 call, which explicitly aims to foster reusable, multi-scale, virtual human twin systems. The SMASH-HCM research toolbox facilitates that future research can be done on the algorithms and APIs with self-provided data.
- Long-term sustainability: Capturing research-specific functional requirements such as containerised deployment, batch simulation capabilities, modifiable model parameters, and open APIs and providing initial toolbox components support the maintainability and continued usage of the SMASH-HCM research toolbox beyond the project's lifecycle.
- Compliance and separation of concerns: While the clinical product must comply with strict regulatory frameworks (e.g., MDR), researchers require sandboxed access to experimental features, simulation engines, synthetic cohorts, and diagnostic algorithms. This deliverable enables such separation by planning functional requirements for a distinct research-oriented configuration of the SMAS-HCM research toolbox.

Accordingly, D7.6 consolidates profound desk-research insights and input from SMASH-HCM researchers. This process was twofold: first, a structured desk-research phase primarily using PubMed and selected regulatory documents to map scientific and technical requirements; second, a validation phase through two dedicated focus groups and a set of bilateral discussions with SMASH-HCM researchers. These exchanges ensured that the requirements are not only theoretically grounded but also reflect practical needs from those working with hypertrophic cardiomyopathy data and models. In this sense, the requirements formulated here are based on contributions from different project beneficiaries, expert discussions, and analysis of relevant scientific and regulatory frame works. They cover functional workflow support for researchers, technical and architectural aspects, performance and scalability expectations, as well as the legal and ethical dimensions of handling sensitive patient data. It focuses on research-enabling functionalities such as standardised data in- and export, scalability for high-performance computing, interoperability with external platforms such as e.g. EDITH or InSilicoTrials.

2 SMASH-HCM as a research tool

Definition: "The SMASH-HCM research toolbox is a modular and flexible collection of computational models, algorithms, and application programming interfaces (APIs) developed within the SMASH-HCM project. It is designed as an open, extensible toolbox rather than a monolithic application, enabling clinicians and researchers to access, combine, and deploy components according to their specific needs."

The toolbox serves as both a methodological resource and an infrastructure for translational research in hypertrophic cardiomyopathy. Its design is guided by three core principles: flexibility, accessibility, and sustainability.

Flexibility and Scope

- The toolbox should provide a set of interoperable components, including disease models, computational algorithms, and APIs.
- Researchers can selectively integrate these components into their own environments, assemble custom pipelines, and benchmark external models against those developed in SMASH-HCM.
- Use cases span clinical and research domains, from per-patient decision support to population-level analyses and exploratory data science.

Integration and Interfaces

- Models and algorithms should be containerised (e.g., via Docker) to support reproducibility, portability, and straightforward test-integration with medical devices or research platforms.
- Future integration with European initiatives such as the Virtual Human Twins (VHT) Initiative or the Ecosystem for Digital Twins in Healthcare platform (EDITH) is encouraged, offering a pathway to long-term interoperability.

Documentation and Accessibility

- To support open science, the toolbox should be hosted on repositories (initially private, with the potential to become public), ensuring transparency, reuse, and accessibility.
- Documentation should accompany all software components, covering both technical interfaces and example use cases.
- Open licensing should be pursued wherever possible. Restrictions are expected e.g. where patient data or sensitive intellectual property considerations apply.

Data and Privacy Considerations

- While open patient and study data sharing may be limited due to privacy regulations, the toolbox may include anonymised example datasets for demonstration and benchmarking purposes.
- Programmatic access to data outputs from SMASH-HCM work packages is encouraged, respecting legal and ethical requirements.

Sustainability and Future Directions

- Long-term sustainability is recognised as a challenge. Potential strategies include integration with established European platforms (VHT, EDITH), institutional hosting of repositories, and further funding or commercialisation initiatives.
- Dissemination through scientific publications, repositories (e.g., GitHub, Zenodo), and conference presentations can ensure visibility and uptake beyond the project.

In summary, the SMASH-HCM research toolbox is not a single software tool but a structured collection of well-documented, containerised, and openly accessible components. It is intended to support both

clinical and research applications, facilitate custom pipeline creation, and provide a sustainable foundation for future integration into European health research infrastructures.

Use-Cases: "The SMASH-HCM research toolbox should be designed to support a spectrum of academic, clinical, and industrial research activities. These activities extend beyond the constraints of a regulated clinical product and take full advantage of the research toolbox's modular digital twin architecture, algorithmic flexibility, and data interoperability. The use cases outlined below describe potential research scenarios anticipated beyond the project lifetime."

1) **In-silico Trials and Preclinical Stratification:** Researchers conducting early-stage therapeutic development or mechanistic investigations can make use of tools to simulate patient cohorts, explore genotype–phenotype relationships, or test drug responses virtually. The SMASH-HCM research tools may allow for:
   - Stratification of virtual patient populations based on molecular, functional, or imaging-derived biomarkers.
   - In-silico simulation of treatment scenarios, such as pharmacological or other interventions or device parameter variations.
   - Batch execution of digital twin simulations for (digital) biomarker discovery, endpoint modelling, and hypothesis testing.
2) **Algorithm Development and Validation:** The research subversion of the SMASH-HCM solution should allow future data scientists and researchers to plug in custom algorithms, test performance across synthetic or retrospective datasets, and assess generalisability across digital cohorts. Use cases include for instance training and validation of machine learning models using labelled outputs from the SMASH-HCM pipeline; or prototyping of new clinical decision rules, stratification thresholds, or predictive indicators.
3) **Model Integration and Multi-Scale Simulation:** The toolbox supports users aiming to integrate new computational models at various biological scales. This use case is particularly relevant for consortia members and external collaborators with interest in extending the research toolbox toward multi-organ simulations (e.g., cardiometabolic coupling, exercise physiology) and in studying the propagation of cellular or molecular abnormalities to patient-level outcomes.
4) **Educational and Academic Use:** Universities and teaching hospitals can adopt the research tool to illustrate digital twin technology, train students in computational cardiology, or teach reproducible modelling workflows. Potential uses include:
   - Course integration in biomedical engineering, data science, or medicine curricula.
   - Supervised projects for MSc or PhD students exploring cardiovascular modelling or in-silico digital-twin methods.
   - Ethical demonstrators for public engagement or stakeholder training.
5) **Translational and Industrial Research:** Pharmaceutical and MedTech partners may use the SMASH-HCM research toolbox as a sandbox for translational innovation. This could include:
   - Using digital twins to refine patient inclusion criteria for preclinical and early-phase studies.
   - Evaluating how device parameters affect physiological indicators within digital patient populations.
   - Informing the development of companion diagnostics or personalised treatment pathways.
6) **Regulatory Science and Standards Development:** Finally, the research tool can support collaborations with regulators or standardisation bodies to interoperability with European infrastructures such as EDITH. This may involve benchmarking model explainability and robustness across different conditions; generating annotated datasets or simulation logs for algorithm auditing; or contributing to open standards for FAIR and transparent digital twin reporting.

3 System Requirements

This chapter describes what the research toolbox should be capable of, and what functional, technical, performance and legal/ethical features would be needed to achieve this. For each requirement, an indication based on researcher review, whether this is achievable within the scope of the SMASH-HCM project ("Implementation planned in SMASH-HCM", "Development after SMASH-HCM has ended") and if it is overall desirable (categories: "not needed", "useful", "required"). The system requirements described in this chapter are the result of a structured researcher consultation process. Two focus groups were organised, involving active participation of researchers and technical experts from multiple SMASH-HCM beneficiary organisations (InSilicoTrials, EMP, TAU, PHARM, and clinical partners). In total, over ten researchers contributed actively, providing feedback on the functional, technical, performance, and legal dimensions of the toolbox. These discussions were complemented by bilateral talks to clarify open questions and to ensure coverage of domain-specific perspectives. The outcome was systematically mapped into the requirement tables presented below. Cells in the tables were colour-coded (green) based on a majority voting procedure: if most participating researchers agreed that a requirement was "required" and feasible, the cell was marked accordingly. This methodological approach provided a transparent and democratic way to consolidate individual inputs into collective system requirements

3.1 Functional Requirements for Clinical Research Workflow Support

To be a valuable research asset, the SMASH-HCM research toolbox must seamlessly fit into clinical research workflows for HCM. This means supporting researchers at every step, from study design and data collection, through analysis and interpretation, to collaboration and publication. In current practice, HCM researchers face laborious data management, as they must extract relevant patient data from electronic health records for studies, and face difficulties in pooling data across hospitals due to heterogeneity and privacy constraints, and challenges in reproducing analyses. A digital research toolbox that directly interfaces with clinical data and provides built-in analytical tools can dramatically improve efficiency. For example, instead of manually querying multiple hospital databases for HCM patients meeting certain criteria, a researcher could use a tool from the SMASH-HCM researcher toolbox to identify a virtual cohort and automatically retrieve their de-identified profiles for analysis. Additionally, modern research often requires iterative experimentation, like tuning models and running simulations. The research toolbox should thus facilitate this iterative cycle by providing an interactive environment with minimal bureaucratic hurdles.

One major challenge is ensuring the research tool does not disrupt clinical operations but rather complements them. Researchers may want to use retrospective clinical data and prospective data from ongoing studies. The research toolbox should connect to hospital electronic health records or data warehouses in a read-only, secure fashion to fetch updated data on consented patients, while also allowing upload of external research datasets (e.g., omics data from lab experiments). Data access must be governed properly and only authorized research users for approved purposes may gain access. Another challenge would be an integration with existing clinical trial management systems. For instance, if a multi-center HCM trial is underway, the research toolkit should be able to ingest trial data such as patient characteristics, endpoints) in near real-time to update digital twin simulations or to stratify patients as the trial progresses.

The following functional requirements are suggested for a research toolbox to support clinical research workflows:

Study Cohort Identification

The toolbox shall enable researchers to **identify and extract cohorts** of HCM patients from multi-center clinical databases based on complex criteria (e.g., genotype, age, symptoms) in a user-friendly manner. It should interface with electronic health records (EHRs) or research registries to pull de-identified participant data for those who meet inclusion criteria. This supports faster cohort assembly and ensures researchers have relevant patient data at hand for analysis.[1]

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Integration with EHR Systems

The system shall connect to hospital EHR or clinical data warehouses via standardized protocols (FHIR APIs or similar) to **automatically import relevant clinical data** (diagnoses, medications, imaging results, lab values, etc.) for research, subject to patient consent and IRB approvals. By using shared standards like HL7 FHIR or OMOP common data models, this integration facilitates interoperability across sites.[2] The research toolbox's querying mechanism should map common research questions to EHR data elements, for example: "retrieve all echocardiography reports for this cohort in the last 5 years".

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended
			SMASH-HCM may have one or more standard protocols implemented in the research toolbox	Full-fledged EMR integration is outside of the scope of SMASH-HCM

[1] See Maxwell, Lauren et al. "FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries." *The Lancet Digital Health*, Volume 5, Issue 10, e712 - e736. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00129-2/fulltext

[2] See Maxwell, Lauren et al. "FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries." *The Lancet Digital Health*, Volume 5, Issue 10, e712 - e736. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00129-2/fulltext

Research Data Ingestion

Beyond EHR data, the research toolbox shall allow ingestion of external research data. Researchers must be able to upload datasets such as genomic sequences, in vitro experimental results, or imaging files obtained outside the hospital environment. The toolbox should provide templates for metadata (documenting how and when data was obtained) to ensure all imported data is well-annotated according to FAIR principles.[3]

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended
			Some data import functionality will be available	

Study Protocol Management

The research toolbox could provide tools to design and document research protocols. This could include a module to define the aims, endpoints, and analysis plan for a digital twin simulation study, like a virtual clinical trial. While not strictly necessary for the toolbox, such a feature would be desirable for keeping all research context in one place. For example, a researcher could outline a "Virtual trial of Drug X on 100 HCM digital twins" within the research toolbox, then execute the simulations and track outcomes.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Analytical Workflows and Notebooks

The solution shall support interactive analysis workflows so researchers can perform data cleaning, statistical analysis, and visualization without needing to constantly export data. This can be achieved by integrating notebook environments (e.g., Jupyter) or built-in analytics dashboards. The aim is to bring the analysis to the data, improving data security and efficiency. Researchers should be able to write and execute code (R, Python, MATLAB, etc.) on the research toolbox against the HCM datasets and digital twin outputs, enabling exploratory analyses and custom model development.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended
	The toolbox could provide interfaces to data and algorithms (via API)			

[3] See Maxwell, Lauren et al. "FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries." *The Lancet Digital Health*, Volume 5, Issue 10, e712 - e736. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00129-2/fulltext

Audit Trail for Research Activities

Every data access, analysis run, and model simulation in the research toolbox shall be logged in an audit trail. This is crucial for later verification of results and for compliance with good research practice. The log should record who accessed what data or ran which simulation, and any changes made to data or configuration. Such auditability aligns with the concept of a "medical algorithmic audit" proposed for clinical AI systems, repurposed here for research integrity. It ensures that published findings can be traced back and verified, which is increasingly demanded by journals and regulators.[4]

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Workflow Automation

It is desirable that the research toolbox supports automation of repetitive tasks in the research workflow. For instance, if a researcher needs to rerun a particular analysis every time new patient data arrives, the system could allow scheduling of that job or triggering it automatically. This ensures that the digital twin models and risk predictions for a cohort are kept up-to-date as fresh data comes in, like new follow-up visits of patients.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Ease of Use and Training

The research toolbox must be usable by clinical researchers who may not have advanced programming skills. A well-designed GUI for cohort selection and data visualization is essential (though coding interfaces can be available for data scientists). Adequate training material and user support should be provided, as the user base will be academics and clinicians. Lowering the barrier to entry encourages broader adoption in the research community.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Compliance with Research Governance

The research toolbox should facilitate compliance with research regulations. For example, it should enforce that only pseudonymized and de-identified data is used in research mode, and that any re-identification (if needed for clinical correlation) follows proper governance. It also should help in maintaining patient consent records, like marking if a patient has consented to their data being used in certain research studies. This ensures that data usage stays within ethical bounds, which is a concern in any health research tool.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

[4] See Corral-Acero, J., Margara, F., Marciniak, M., Rodero, C., Loncaric, F., Feng, Y., Gilbert, A., Fernandes, J. F., Bukhari, H. A., Wajdan, A., Martinez, M. V., Santos, M. S., Shamohammdi, M., Luo, H., Westphal, P., Leeson, P., DiAchille, P., Gurev, V., Mayr, M., Geris, L., ... Lamata, P. (2020). The 'Digital Twin' to enable the vision of precision cardiology. *European heart journal*, *41*(48), 4556–4564. https://doi.org/10.1093/eurheartj/ehaa159

Interoperability with Trial Systems

As a desirable feature, the research toolbox could interoperate with existing clinical trial management systems (CTMS). For instance, if a clinical trial is registered and managed in a CTMS, basic info (enrolment, outcomes) could be pulled into SMASH-HCM for analysis. Conversely, results from digital twin simulations that predict patient risk could inform stratification in a trial, with proper blinding. Standards for such data exchange, like using trial IDs or data export and import formats, should be followed. This would align with calls for more integrated data systems in research, as noted during COVID-19 where linking various data resources was beneficial.[5]

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended
It could be useful in view that the research toolbox provides interfaces allowing an integration with clinical trial systems.				

Implementation Guidance

To implement workflow requirements, early integration with hospital IT is needed. The use of common data models (OMOP CDM or HL7 FHIR) will simplify EHR connectivity. For example, mapping HCM patient data to a Patient Digital Twin Profile. Experts recommend expanding efforts to link EHRs by shared standards like FHIR or OMOP, and applying standardized terminologies (SNOMED CT, LOINC) for research data. The SMASH-HCM research tool development team should collaborate with data managers at the SMASH-HCM clinical site to set up data pipelines that regularly push updates to the research database. On the user interface side, iterative design with input from researchers will ensure the tools align with researchers' habits. For example, a cardiologist-scientist might want to query data via a form interface, whereas a data scientist might prefer writing a query or script. The research toolbox could cater to both. Importantly, given that researchers will use real patient data, privacy safeguards and user authentications must be rigorous. In summary, the SMASH-HCM research tool should *embed itself into the researcher's workflow*, reducing manual effort and enabling focus on interpretation and discovery.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended
			Documentation on how to use/implement the toolbox will be available with D7.9	

[5] See Maxwell, Lauren et al. "FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries." *The Lancet Digital Health*, Volume 5, Issue 10, e712 - e736. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00129-2/fulltext

3.2 Technical Architecture and Implementation Requirements

This section describes how the research tool should be built to support the aforementioned functional requirements.

Modular Model Architecture

Desirably, the modelling architecture should be modular and extensible. Researchers might develop new model components (say, a new module for metabolic remodelling in HCM) and should be able to plug it into the twin framework. A modular design with defined API interfaces would allow innovation and extension. While not absolutely required for initial functionality, this is important for a research tool as it will evolve with scientific progress. It encourages community contributions and adaptation for related diseases in the future.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Cloud-Based or Cluster Infrastructure

The research toolbox shall be built on a scalable cloud or cluster infrastructure. This means it should run on an environment where computing resource can be allocated dynamically. The architecture should separate the web/front-end from the back-end compute. The infrastructure should support horizontal scaling (adding more compute nodes for more tasks/users) and possibly vertical scaling (using bigger nodes for very large tasks). By using containerization and orchestration (e.g., Kubernetes, Slurm on HPC), the research toolbox can manage and distribute jobs efficiently. This approach aligns with modern scientific platforms which use cloud elasticity to handle variable workloads.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended
			By using containers, the components can be easily deployed in a cloud-service	

Interoperability with Modality-specific Tools

It would be beneficial if the research toolbox can plug into existing specialized tools for each modality. For instance, link out to a genomics browser for detailed sequence view, or send an imaging study to a PACS viewer or Osirix DICOM viewer for radiologists. This interoperability ensures researchers are not constrained if they need more advanced analysis than the research toolbox provides natively. Using standard protocols like DICOMweb for images, FHIR Genomics for variant data, etc. could help achieve this.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Backup, Archiving, and Preservation

In line with good data management, and in an ideal scenario, a research toolbox should regularly back up data and provide archival solutions for older or completed studies. Research data might need to be preserved for years (for reproducibility or future reuse). The refore, the SMASH-HCM research toolbox should implement retention policies, such as moving raw data from completed studies to cheaper long-term storage after a period but continuing to allow the metadata to be searchable. This ensures that even after the project or as years pass, the data remains available. However, for the purpose of this research project and the anticipated use-case of the research toolbox after SMASH-HCM has ended, the focus group decided that regular data backups and archiving was not needed.

Not needed	**Useful**	**Required**	**Implementation planned in SMASH-HCM**	**Development after SMASH-HCM has ended**

High-Availability & Load Balancing

In an ideal scenario, the research toolbox could be deployed in a high-availability setup, e.g., multiple instances behind a load balancer, so that even if one instance goes down, others continue serving (zero downtime). This is especially relevant for the front-end/API services. For research use, a brief downtime might be tolerable, so it would be a desirable rather than an essential requirement.

Not needed	**Useful**	**Required**	**Implementation planned in SMASH-HCM**	**Development after SMASH-HCM has ended**

Maintainability and Updates

The research toolbox's infrastructure and software should be maintainable, meaning it can be updated and patched with minimal downtime. Using containerization new versions of services can be deployed gradually (blue-green deployments, etc.). Regular updates are needed not just for security purposes but also for improvements such as new model versions and new features. The architecture should allow module-wise updates (e.g., update the AI model component without overhauling the entire system). Documentation for system admins is needed (how to scale up resources, how to recover from failures, etc.).

Not needed	**Useful**	**Required**	**Implementation planned in SMASH-HCM**	**Development after SMASH-HCM has ended**

User Interface for Multi-modal Data

A user-friendly UI is desirable to navigate multi-modal data. This could be a dashboard per patient showing key genomics such as a list of pathogenic variants, key imaging measurements, and recent PROM scores. Such a dashboard would help researchers qualitatively understand each case and pick out anomalies or patterns that merit further investigation. It would also aid in presentations or teaching and visualizing a digital patient.

Not needed	**Useful**	**Required**	**Implementation planned in SMASH-HCM**	**Development after SMASH-HCM has ended**

3.3 Performance, scalability, availability, usability Requirements

This section describes what requirements are needed for the research toolbox's performance.

Scalable Storage and Processing for Large Modalities and Computing-Intensive tasks

Imaging and genomics are data-heavy, and some of the algorithms computation intensive, so the research toolbox's infrastructure must be capable of storing large images and performing heavy computations. For example, analysing cardiac MRIs for 100 patients might require significant CPU/GPU time. Thus, scalability and performance are critical. The system should use efficient data formats like compressed Digital Imaging and Communications in Medicine (DICOM) or Neuroimaging Informatics Technology Initiative (NIfTI) for images, compression for genomic data, and possibly pre-compute key results, like image segmentations, to avoid re-running costly operations repeatedly. Caching mechanisms for frequently accessed multi-modal data can improve responsiveness. The non-functional requirement is that adding more images or sequences does not slow down the research toolbox. Hence cloud scaling and parallel processing are needed.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Model Safety (for Clinical Translation)

Although used in research, eventually these models might inform decisions in clinical practice. Non-functionally, it is important that the models avoid unsafe or unrealistic outputs without flagging them. For example, if a simulation yields an outrageous result, as in negative blood pressure results or 100% risk, the system should catch that and warn the user that the model is outside valid bounds. Essentially, model output sanity checks should be in place. This protects against overinterpreting model extrapolations. It is part of responsible research to know when the model might be breaking down due to inputs outside training range.[6]

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

User Control and Transparency

Desirable is a level of **user control and transparency over the model**. Researchers might want to tweak model assumptions or see the equations. The research toolbox should not be a "black box." It could provide documentation of the model (equations, sources), and possibly an interface to adjust advanced settings (like using a different action potential models or toggling on/off certain physics). This caters to the academic audience who might be interested in the model itself, not just its output, and fosters trust when they can inspect it.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

[6] See Corral-Acero, J., Margara, F., Marciniak, M., Rodero, C., Loncaric, F., Feng, Y., Gilbert, A., Fernandes, J. F., Bukhari, H. A., Wajdan, A., Martinez, M. V., Santos, M. S., Shamohammdi, M., Luo, H., Westphal, P., Leeson, P., DiAchille, P., Gurev, V., Mayr, M., Geris, L., ... Lamata, P. (2020). The 'Digital Twin' to enable the vision of precision cardiology. *European heart journal*, *41*(48), 4556–4564. https://doi.org/10.1093/eurheartj/ehaa159

Model Explainability Interface

The research toolbox shall provide explainability tools for AI/ML models integrated into the system. For every prediction or output given by an AI component (be it a risk score, a classification, or regression), the user should have the option to "explain this result." For example, if an AI model says a patient has a 20% 5-year risk of sudden cardiac death, the system could display the top contributing factors: "family history of sudden cardiac death" contributed 5%, "extensive Late Gadolinium Enhancement scar on MRI" contributed 3%, etc., perhaps via a visual bar chart or textual summary. In imaging contexts, highlight regions of the image that led to a classification (heatmap overlay).[7] In time-series (wearable data), identify which segments or features influenced the result. The requirement is that these explanations are readily accessible and understandable to the researcher and specifically a clinician. This might involve integrating known techniques: *saliency maps* for images, *SHAP values* for tabular data, *attention weight visualization* for sequence models, etc. The aim is to demystify the "why" behind model outputs, thereby aiding scientific insight. For digital twin simulations (mechanistic models), explainability is less about black-box math and more about output interpretation – but we should similarly provide clarity (e.g., if the simulation predicts an arrhythmia, highlight which tissue region or parameter caused it).

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Provenance Tracking and Versioning

The research toolbox shall implement comprehensive provenance tracking for data and models. This means any result can be traced back to its origins: which raw data, which pre-processing, which model version, which parameter settings. For instance, if a graph is generated in a publication from the research toolbox, one should be able to retrieve an exact "analysis recipe" or script that produced it, with references to data inputs (perhaps via dataset DOIs) and model IDs. All ML models should be versioned (with unique identifiers and metadata like training date, algorithm, training dataset, performance metrics). If a model is updated or retrained, it gets a new version but the old one remains accessible for reproducibility or audit. Similarly, when datasets are updated (e.g., a correction in a data entry), versioning ensures previous analyses can still be repeated on the earlier snapshot if needed. This is akin to maintaining a code repository but for *analysis pipelines*. Researchers should be able to "publish" their pipeline configurations or share them internally. This requirement goes beyond logging; it's about structured, queryable records of how each result was derived. It aligns with recommendations to have machine-readable metadata and checklists for data availability and reproducibility.[8]

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

[7] See Li, X., Loscalzo, J., Mahmud, A.K.M.F. *et al.* Digital twins as global learning health and disease models for preventive and personalized medicine. *Genome Med* **17**, 11 (2025). https://doi.org/10.1186/s13073-025-01435-7

[8] See Maxwell, Lauren et al. "FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries." *The Lancet Digital Health*, Volume 5, Issue 10, e712 - e736. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00129-2/fulltext

Reproducible Analysis Environments

The research toolbox shall provide reproducible computational environments for running analyses. For example, if a researcher wrote a Python script in the research toolbox's notebook to analyse data, the environment (library versions, OS) should be consistent or encapsulated. Possibly use container technology or environment management under the hood such that anyone re-running that script on the research toolbox uses identical package versions. Additionally, the research toolbox should allow exporting the entire analysis (code + environment + data snapshot) in a form that can be run outside (for transparency or journal supplemental material) – for instance, exporting a docker container or a workflow file (e.g. CWL). This ensures that even outside the research toolbox, reproducibility is possible, which is often needed for peer review or external validation.[9] The concept of a "reproducibility package" for each result could be implemented.

Not needed	**Useful**	Required	**Implementation planned in SMASH-HCM**	**Development after SMASH-HCM has ended**

Audit Trails for Model Development

Whenever a user trains or modifies an AI model from the research toolbox, an audit trail shall record the process. This includes data used for training (with version), hyperparameters, training duration, and performance metrics on validation/test data. If later that model is used in analysis or suggestion, one can audit how it was created and how it was tested. This addresses concerns that AI can learn shortcuts or biases; an auditor could inspect, for example, if a model was validated on diverse subgroups or if certain error modes were observed.[10] The research toolbox could facilitate "algorithmic auditing" by providing modules to test a trained model on stratified subgroups or generate adversarial scenarios. But minimally, the training log and evaluation should be preserved for inspection.

Not needed	Useful	**Required**	**Implementation planned in SMASH-HCM**	**Development after SMASH-HCM has ended**

Monitoring Model Behaviour for Drift

Over time, as new data are added or as the patient population changes, machine learning models might drift, and their performance may degrade, or their predictions ratio might change. The research toolbox could include a feature to monitor model performance over time or across subgroups as new validation data comes in. If an explainable AI suddenly starts giving odd attributions or errors for a subset, maybe due to drift, that can be flagged. This is more relevant in live deployment, but in research, if the cohort composition changes, it is good to detect if models need retraining. However, this feature is not critical for initial research phases.

Not needed	**Useful**	Required	**Implementation planned in SMASH-HCM**	**Development after SMASH-HCM has ended**

[9] See Corral-Acero, J., Margara, F., Marciniak, M., Rodero, C., Loncaric, F., Feng, Y., Gilbert, A., Fernandes, J. F., Bukhari, H. A., Wajdan, A., Martinez, M. V., Santos, M. S., Shamohammdi, M., Luo, H., Westphal, P., Leeson, P., DiAchille, P., Gurev, V., Mayr, M., Geris, L., ... Lamata, P. (2020). The 'Digital Twin' to enable the vision of precision cardiology. *European heart journal*, *41*(48), 4556–4564. https://doi.org/10.1093/eurheartj/ehaa159

[10] Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health. 2022 May;4(5):e384-e397. doi: 10.1016/S2589-7500(22)00003-6. Epub 2022 Apr 5. Erratum in: Lancet Digit Health. 2022 Jun;4(6):e405. doi: 10.1016/S2589-7500(22)00089-9. PMID: 35396183.

User Education and Interpretability

Users of the research toolbox need to understand the explainability outputs. The toolbox should include guidance and documentation (e.g. in SMASH-HCM project deliverables D7.8 and particularly D7.9) on interpreting explainability visualizations and metrics. For example, an explanation dashboard should have a legend and perhaps text that interprets "the following factors increased the risk prediction." This helps avoid misinterpretation. This further allows to educate researchers on pitfalls, like not over-trusting partial explanations. In essence, incorporating some training or help within the UI for these advanced features might be useful.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

Integration of Notebooks and Narrative

Many researchers use Jupyter notebooks or R markdown to weave narrative with code and results for interpretability to others. The research toolbox could integrate such literate programming tools so that the analysis, results, explanations, and discussion can live together. That narrative can then be exported as a PDF or HTML and serves as a lab notebook record (for audit) and supplementary material (for reproducibility). Encouraging this practice yields clearer analyses and easier external understanding. This is not strictly required within the SMASH-HCM project lifetime, since people can do this off-line, but integration would smooth the workflow and thus our SMASH-HCM experts flagged this as "required after the project as ended".

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

3.4 Legal and Ethical Requirements

The SMASH-HCM research toolbox deals with sensitive patient data that could influence health decisions. It is therefore essential that the research toolbox complies with all relevant regulatory frameworks, follows ethical standards, and protects patient privacy. If the research toolbox is developed with legal and ethical regulations in mind, it will later allow for a smoother translation of research findings into clinical practice or trials.

Regulatory compliance involves data protection laws, like the General Data Protection Regulation (GDPR) in the EU, as well as research-specific regulations. For example, if the tool were later to be used to support decisions on patient management, it might border on being a medical device, which would invoke medical device regulations like the EU MDR. In a research context, the tool is not providing direct care decisions, but it should be designed with the strictest standards in mind in case of future deployment.

Guaranteeing Privacy is of utmost importance in this context, as HCM patient data includes genetic information (highly identifiable and sensitive), clinical records, etc. The GDPR classifies health and genetic data as special categories that require strong safeguards and typically explicit consent for research use.[11] Each patient whose data enter the research toolbox has informed consent attached to it (this is handled in WP8 ethical approvals[12]), but the research toolbox must enforce those consent restrictions and ensure confidentiality. A quote from a Genome Medicine article highlights a key ethical question: "can a patient be asked to share detailed information from her digital twin as a resource for

[11] See Art. 9 GDPR https://gdpr-info.eu/art-9-gdpr/

[12] See Grant Agreement - GAP-101137115

treating similar patients or for research?".[13] The SMASH-HCM research toolbox is essentially doing that – using patient data (via the twin) for broader research. This must be addressed by integrated solutions covering ethics, data security, and regulatory compliance.[13]

Ethical considerations also include transparency. Patients and the public should have clarity on how their data and digital twin are used. They also include benefit sharing, to ensure that research benefits will eventually help patients, and oversight, so that ethical boards can review studies that have been done using the research toolbox. The research toolbox should enable ethical practices like audit trails for accountability and accommodate things like withdrawal of consent.

Functional Requirements (Regulatory & Privacy):

Privacy-Preserving Data Analysis Methods

Where possible, the research toolbox should implement privacy-preserving analysis techniques. One example is federated learning, which allows building machine learning models across multiple sites' data without centralizing raw data. If the SMASH-HCM research toolbox could involve multiple hospitals that cannot share data outright, the research toolbox could send models to train at each site and then aggregate their outputs or parameters. The requirement is to support such methods, which might include federated learning, secure multi-party computation, or data anonymization for certain outputs. This would significantly enhance the reach of the research toolbox. For example, if one hospital's data cannot leave its premises, the research toolbox's federated module can still include that data in modelling by training locally there and only sharing model weights which could be a soft-workaround that is less sensitive. Sources recommend addressing root causes of data silos like legal concerns and using solutions like federated analysis to protect privacy.[14][15] If the SMASH-HCM research toolbox actively does this, it will be a strong compliance measure and potentially attract participation from centres that otherwise are hesitant to share their patients' data.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

[13] Li, X., Loscalzo, J., Mahmud, A.K.M.F. *et al.* Digital twins as global learning health and disease models for preventive and personalized medicine. *Genome Med* 17, 11 (2025). https://doi.org/10.1186/s13073-025-01435-7

[14] See Maxwell, Lauren et al. "FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries." *The Lancet Digital Health*, Volume 5, Issue 10, e712 - e736. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00129-2/fulltext

[15] See Li, X., Loscalzo, J., Mahmud, A.K.M.F. *et al.* Digital twins as global learning health and disease models for preventive and personalized medicine. *Genome Med* 17, 11 (2025). https://doi.org/10.1186/s13073-025-01435-7

Ethical AI Practices

Since the research toolbox includes AI/ML models, it should follow ethical AI practices. This can be non-functional in the sense of guidelines or constraints, like bias mitigation strategies, fairness checks on models, and explainability as part of core design. The research toolbox might incorporate bias detection tools for checking if a predictive model's performance differs by gender or ethnicity (if such data is available or can be proxied). AI should be transparent and fair. While not a direct regulatory requirement yet, ethical regulations are being considered. For example, the EU has proposed an AI Act in 2024 that classifies health AI as high-risk, meaning that it would need to comply with specific requirements.[16] By implementing these practices now, the research toolbox ensures that any AI developed on it can meet future regulatory scrutiny.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended
		Support for ethical AI practices should be provided		

International Regulatory Adaptability

If the research toolbox eventually sees use beyond the EU it should be adaptable to other regulatory regimes as well (i.e., HIPAA in the USA). This is more a design philosophy to keep the research toolbox modular. For example, if at some point it is necessary to apply a "Safe Harbor" de-identification standard for HIPPA,[17] or to comply with specific national requirements, like Germany's data localisation, it would be possible. This, however is not a crucial requirement for the immediate project, but rather good for its longevity.

Not needed	Useful	Required	Implementation planned in SMASH-HCM	Development after SMASH-HCM has ended

[16] See https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en

[17] The HIPPA „Safe Harbor" de-identification is a method used to remove specific data elements from protected health information to render it non-identifiable.

4 Conclusion

This deliverable D7.6 has presented a consolidated set of requirements for the SMASH-HCM solution in its function as a research toolbox. It reflects the intention of the project to provide not only a clinical product but also a sustainable scientific toolbox that can continue to generate value in research, training, and innovation. By explicitly separating clinical and research perspectives, the deliverable enables a clear distinction between what must be tightly regulated for patient care (D7.1 and D7.2) and what can be more open, flexible, and exploratory for scientific progress (this deliverable D7.6).

The requirements formulated here are based on contributions from different project beneficiaries, expert discussions, and analysis of relevant scientific and regulatory frameworks. They cover functional workflow support for researchers, technical and architectural aspects, performance and scalability expectations, as well as the legal and ethical dimensions of handling sensitive patient data. This wide scope is necessary, because the research toolbox will only be valuable if it addresses the full chain of research activities, from accessing and preparing data to running simulations, documenting workflows, and disseminating results.

A recurring theme throughout the requirements is modularity. Instead of aiming for a single software, the project proposes a collection of containerised components, documented APIs, and well-defined interfaces. This modularity allows researchers to pick and choose the elements that are relevant for their work, to integrate them into their own computing environments, and to extend the research toolbox with new models or algorithms. In this way, SMASH-HCM contributes to a research culture where digital twin technology becomes reusable, interoperable, and adaptable across different scientific communities.

Another central element is sustainability. The deliverable acknowledges that many research projects create promising software prototypes, but after the end of funding these tools are often not maintained. By documenting requirements for sustainability from the outset, D7.6 aims to avoid this typical risk. Strategies such as integration with European platforms (e.g. Virtual Human Twins, EDITH), hosting of code and models in open repositories, and alignment with open science practices are identified to keeping the toolbox alive after the project ends. At the same time, the document does not ignore the challenges: long-term hosting, updating, and support will require institutional backing or new funding sources, and therefore follow-up planning is necessary beyond this deliverable.

Ethical and legal safeguards are another important part of the requirements. The SMASH-HCM research toolbox may deal with data that is sensitive and protected under GDPR and other regulations. The deliverable stresses the need for strong privacy-preserving mechanisms and transparent audit trails. It also addresses emerging concerns in artificial intelligence, such as explainability, bias detection, and fairness. These measures not only protect patients and research participants but also build trust among users and stakeholders that the toolbox can be used responsibly.

In summary, this deliverable provides a roadmap for turning SMASH-HCM into a valuable research toolbox. It is not simply a wish list of features but a structured framework that distinguishes between what is essential, what is desirable, and what may be developed in the longer term. By doing so, it gives guidance to SMASH-HCM developers and researchers alike, and it sets realistic expectations for the project's outputs.

Looking ahead, the requirements documented here may be revisited in Deliverable D7.9, where the consortium will report on which of them were implemented and which remain open. This process ensures accountability and transparency.

5 List of Abbreviations

Abbreviation	Definition
API	Application Programming Interface
CDM	Common Data Model
CTMS	Clinical Trial Management System
DICOM	Digital Imaging and Communications in Medicine
DICOMweb	Web-based protocol for DICOM image exchange
EHR	Electronic Health Record
EDITH	Ecosystem for Digital Twins in Healthcare
EMP	empirica Gesellschaft für Kommunikations- und Technologieforschung mbH
FAIR	Findable, Accessible, Interoperable, Reusable
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HCM	Hypertrophic Cardiomyopathy
HPC	High-Performance Computing
HL7	Health Level Seven (standards organisation)
IRB	Institutional Review Board
LOINC	Logical Observation Identifiers Names and Codes
MDR	Medical Device Regulation (EU)
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NIfTI	Neuroimaging Informatics Technology Initiative (file format)
OMOP	Observational Medical Outcomes Partnership
PACS	Picture Archiving and Communication System
PROM	Patient Reported Outcome Measure
SCD	Sudden Cardiac Death
SHAP	Shapley Additive Explanations (explainability method for ML)
Slurm	Simple Linux Utility for Resource Management (HPC workload manager)
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
UI	User Interface
VHT	Virtual Human Twins (European initiative)
WP	Work Package