DELIVERABLE

Project Acronym: **SMASH-HCM**

Grant Agreement number: **101137115**

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

D6.2 – Initial Evidence-Based Models and Biomarkers

Revision: 1.0

Authors and Contributors	Name (beneficiary): Agakov (Pharmatics), Orchard (Pharmatics), Rossiter (Pharmatics), Ansari (Pharmatics), Chen (Pharmatics), Dang (TAU), Coleman (UOX), van Gils (TAU), Mainardi (POLIMI)			
Responsible Author	(Felix Agakov)		Email	felix@pharmaticsltd.com
	Beneficiary	PHARM	Phone	+447899881575

Project co-funded by the European Commission within HORIZON-HLTH-2023-TOOL-05-03		
Dissemination Level		
PU	Public, fully open	PU
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
0.01	05/06/25	Agakov, Orchard, Rossiter, Ansari, Chen, Dang, van Gils	Pharmatics, TAU	Aggregated and summarised first draft for D6.2 with additional contributions from TAU.
0.02	10/06/25	Agakov, van Gils, Mainardi	Pharmatics, TAU, Polimi	A few additional functions demonstrating various model types in Sec 3.3.3; a few additional references; minor editing.
0.03	12/06/25	Coleman	UOXF	Comments + tracked changes
1.00	16/06/25	Agakov, Long, van Gils	Pharmatics, TAU	Modifications according to the internal comments

Date of delivery	Contractual:	30.06.2025	Actual:	18.06.2025
Status	final **x** /draft			

Abstract (for dissemination)	Deliverable D6.2 describes the initial SMASH-HCM library of evidence-based models of hypertrophic cardiomyopathy (HCM) and related outcomes, aimed at supporting research and clinical decision-making. These models were developed using Pharmatics' MedAI approach, which mines scientific biomedical literature and transforms findings into libraries of predictive tools for science and medicine. The constructed library of predictive models for HCM are accompanied by an automatically generated, browsable programmer's reference manual and technical documentation, with the specifications adhering to the initial requirements outlined in D7.3 and D7.4. As the project advances, these evidence-based models will be integrated with SMASH-HCM datasets. They will serve as *a priori* predictions, features, or initial layers in multilayered architectures, or experts in conditional mixtures for data-driven predictive models, facilitating research and clinical decisions in WP6-7.
Keywords	Evidence-based models, literature-based models, knowledge-based models, API.

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.

Executive Summary

This report, Deliverable D6.2 of the SMASH-HCM Project, details the development of evidence-based predictive models for clinical endpoints and outcomes in hypertrophic cardiomyopathy (HCM), along with common complications like atrial fibrillation, stroke, and heart failure. It partially builds on our previous Deliverable Reports D7.3, "Initial Design Specification for Decision Support Algorithms", and D7.4, "Initial Application Programming Interface for Decision Support Algorithms", focusing on evidence-based model construction. The developed evidence-based functions were programmed and included in the initial application programming interface (API) instance, available locally and on a virtual private server for development and testing purposes within the SMASH-HCM project consortium. The primary goal of the report is to illustrate the evidence-based model development, software development, and the methodological approach for population characterisation of SMASH-HCM patients, which may inform model calibration and refinement for distinct sub-populations. The secondary goal is to present initial results, including function and population characterisation, along with the programmer reference manual and technical specification of a subset of evidence-based SMASH-HCM functions.

Similar to our previous deliverable reports, D7.3 and D7.4, which focused on technical specifications and API developments, this report begins with an overview of the tasks of work package 6, with a particular focus on the possible future integration of evidence-based models across other tasks in the work package. The report then outlines Pharmatics' MedAI methodology for information retrieval and model synthesis, summarising the AI and machine learning approaches while withholding commercially sensitive information and know-how of the development pipeline. This is followed by descriptions of initial approaches for population analysis, aimed at identifying potential sources of population heterogeneity for model refinement and calibration, and the initial steps carried out for constructing a functioning version of the API in Section 2. Initial results, including function and population characterisation and API description, are presented in Section 3. Due to space constraints, technical specifications are given for a subset of 6 HCM functions from a set of 446 functions concerning HCM outcomes and complications extracted from literature and included in the initial API; however, a link to a browsable reference manual for the entire set is provided.

Deliverable Report D6.2 should be viewed as an evolving document, describing the initial API of evidence-based models that will develop over time. The models are envisaged to be refined and validated using SMASH-HCM datasets, with refined and recalibrated versions eventually replacing the initial constructions. The final library of evidence-based models will support risk prediction, biomarker discovery, and clustering across WP6, help with developing the clinical decision support system in WP7, and potentially support the pilot trial in WP8 and other external studies outside SMASH-HCM. This initial construction of evidence-based models is a step towards creating the world's most comprehensive library of statistical models for HCM and related complications, with a built-in capacity for continual updates based on the latest global research.

1 Introduction

Work Package 6 focuses on the integration of knowledge- and data-driven predictive modelling to advance patient stratification and outcome prediction in hypertrophic cardiomyopathy (HCM), using a broad range of clinical data sources and medical information. The work is organised into complementary tasks including the extraction, preprocessing, and analysis of both unstructured and structured clinical data for both population- and patient-level variables, together with statistical information extracted from relevant scientific biomedical publications.

Task 6.1 is dedicated to extracting information from biomedical literature and transforming it into evidence-based statistical predictive models in a reusable form for subsequent tasks. Predictions from these models can be used as feature sets or as knowledge-based structural constraints within algorithms and architectures during further model development, enhancing the data-efficiency, robustness, and interpretability of the developments in downstream tasks targeting patient referrals, prediction, and multi-level risk stratification. The methodology in Task 6.1 builds on an extended version of Pharmatics' MedAI pipeline, which employs state-of-the-art language models, text embeddings, and cascade architectures to systematically identify sources of evidence most relevant to the targeted HCM subpopulations and outcomes, including systematic review, trials, and primary research articles. The pipeline then extracts and synthesises risk scores and statistical predictive models encoded in tables and other quantitative research results from these publications. The models derived by MedAI and subjected to quality control will be made available to project participants as a software library. This deliverable is based on the initial research and development of Task 6.1. Its role in the wider context of WP6, *Data-based Modelling and Integration*, is summarised in Appendix A: Evidence-based Modelling in the Context of WP6.

Central to the developments of Task 6.1 is Pharmatics' biomedical literature-mining system, MedAI – an AI/ML-based information extraction pipeline that forms a foundation for synthesising statistical knowledge-based models from scientific biomedical literature. In the first step, MedAI employs large language model (LLM) embeddings, which convert text into rich numerical vectors capturing semantic meaning. Cascade classifiers are then used, which sequentially filter and identify relevant evidence, focusing on sources relevant to the specified populations, interventions, outcomes, and data dictionaries for accessible datasets. This approach ensures alignment with specific tasks and data requirements for prediction and stratification. From these sources, the system retrieves and harmonises quantitative information such as model coefficients, decision rules, risk scores, and population summary tables, standardising these elements to construct a library of *a priori* predictive models and risk stratification tools. These models are not simply used in isolation; rather, they can provide priors, features, and even architectural constraints directly embedded into the data-driven learning processes of subsequent tasks. For example, in Task 6.2, risk scores and features extracted from literature can help structure and initialise interpretable referral models, improving their generalisability and explainability by grounding them in validated knowledge. In Task 6.3, these same literature-based priors and predictions can support multimodal deep learning methods, acting as features or constraints. This approach can improve data efficiency and robustness, helping to mitigate overfitting in smaller or more complex patient datasets by providing informative starting points instead of relying on random initialisations during model training. In Task 6.4, predictions from literature-derived evidence-based models can serve as inputs for ensembles and multi-level model architectures. This approach ensures closer alignment of clustering and stratification algorithms with established clinical knowledge, potentially enhancing their accuracy, relevance, and clinical uptake. In Task 6.5, MedAI can enhances fairness, robustness, and reliability of predictions while reducing bias by incorporating models trained on diverse global populations.

The development of the HCM instance of MedAI in this Deliverable 6.2 is motivated by the intuition that the inclusion of evidence-based models facilitates multi-source information-based transfer learning, which involves transferring models or features from the source domains (tasks described in

literature) to target domains (such as biomarker discovery and risk stratification in SMASH-HCM), enabling more efficient and robust learning. Specifically, pretrained models synthesised from published HCM research can be weighted, adapted, or incorporated into the regularisation objective based on the similarity between discovery and target populations and their individual or combined performance in the target domain. This facilitates clinical relevance and generalisability of results derived from global populations to new clinical settings in targeted populations. Moreover, MedAI's extensibility and structural explainability, rooted in its reliance on published clinical research, may facilitate clinical and industrial uptake by enabling the tracing and incorporation of new clinical and epidemiological evidence. Also, by refining the predictions of literature-based models in the target domains and datasets, the approach enables automated recalibration and refinement of published findings in specific settings. This has the potential to accelerate model development, enhance robustness of risk prediction and stratification, and enable tailored validation, refinement, and model orchestration across multiple tasks within SMASH-HCM.

2 Methodology Outline

This section outlines Pharmatics' current and planned developments of the MedAI methodology [1] for transforming scientific findings from biomedical literature into predictive tools, with emphasis on the generation and characterisation of statistical predictive models for HCM and its complications. MedAI is an actively developed AI-powered pipeline that systematically scans and processes open-access publications from sources such as PubMed Central and clinical trial repositories. For each researcher-defined condition, the system employs a sequence of specialised steps for extraction, harmonisation, validation, and documentation of predictive models. This methodological overview is followed by TAU's initial methodology for analysing HCM subtypes, aiming to characterise sources of phenotypic heterogeneity within SMASH-HCM datasets and inform the recalibration and refinement of evidence-based models for distinct subpopulations.

2.1 Model Development

The development of the initial library of evidence-based models was guided by the Pharmatics MedAI process described below, which was adapted and enhanced for HCM and its common complications, resulting in the HCM-specific instance of the API, MedAI-HCM. This ongoing effort is aimed at continuously updating and expanding a library of evidence- and knowledge-based statistical predictive models derived from scientific biomedical literature. The aim is to ensure the models remain current and relevant, supporting new discoveries, clinical development, and patient care in HCM, as well as advancing research and development activities throughout the SMASH-HCM project. Unlike fully manual approaches to constructing clinical risk calculators or reproducing scientific research results, many stages of the MedAI pipeline are fully or partially automated, enabling greater scalability and efficiency. In contrast to methods that rely solely on language models or related generative approaches, the MedAI pipeline is focused on robust model construction from extracted evidence, rather than on autoregressive text generation known to be prone to hallucinations and limited reliability. The description below omits commercially sensitive technical details and information.

2.1.1 Function Generation

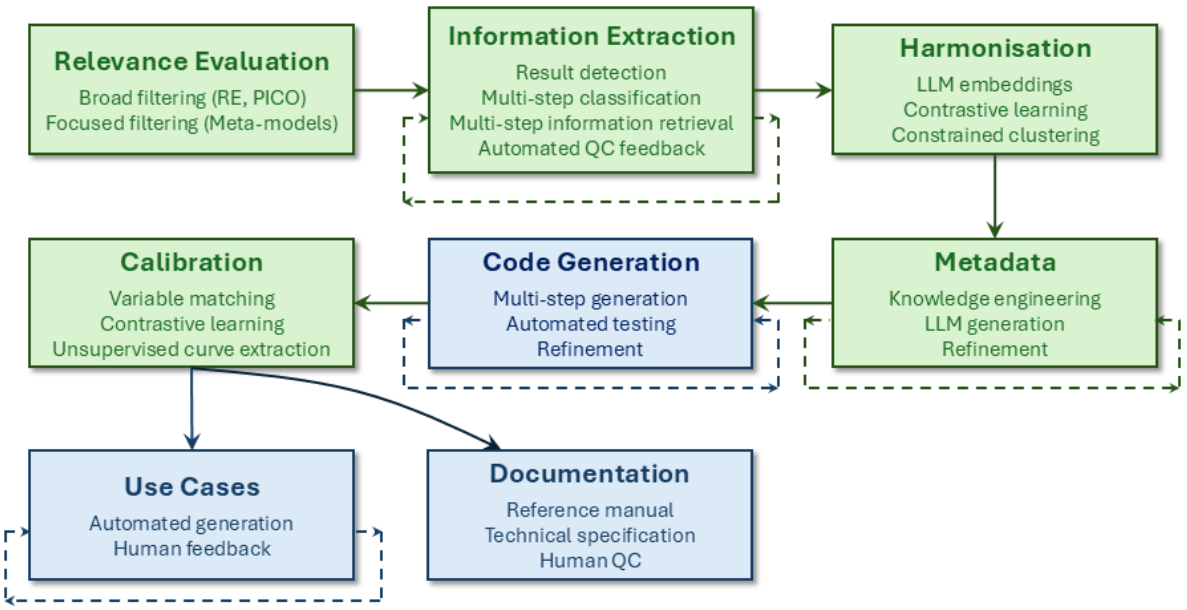

Figure 1 **A schematic illustration of the MedAI evidence-based model development pipeline.**

Function generation deploys a cascade of machine learning methods of increasing granularity and computational cost. The process begins with high-level relevance evaluation applied to millions of

research records from open-access repositories, using ensembles of computationally efficient techniques such as regular expressions, probabilistic string matching, and clustering in the space of lightweight self-supervised LLM embeddings to rapidly screen for preliminary relevance to investigators' queries. This stage typically identifies a subset of 0.1–2% of publications containing results that appear relevant based on population, intervention, and outcome similarity to specific research questions. Once this subset is identified, larger LLM embeddings are deployed within supervised and semi-supervised classification frameworks to refine relevance assessment, detect the presence of predictive models' coefficients or risk assessment scores, and identify specific model types described in scientific publications. These (semi-)supervised methods, although more computationally and financially demanding (due to increased computational costs and the requirement for manually labelled training examples) offer substantial improvements in identifying relevant documents and extracting detailed information about the statistical models, such as standard binary or multi-class classification models or survival regression analyses. The final stages employ semi-automated calibration and quality control procedures, incorporating human oversight to ensure accuracy and consistency in the extracted and synthesised models.

Some of the key steps of the process are summarised on Figure 1 and outlined below.

Relevance Evaluation

The initial step focused on evaluating the relevance of a corpus of approximately nine million full texts of biomedical publications, guided by the PICO (Population, Intervention, Comparison, Outcome) framework established in evidence-based medicine. To identify publications relevant to specific research queries, we first applied an unsupervised pre-screening stage, using regular expression (RE) pattern searches to detect specific terms for the PICO components, followed by probabilistic string matching to capture partial text overlaps and simpler lightweight LLMs to generate vector space embeddings of abstract and result contents for semantic similarity matching. These methods allowed us to rank all publications based on an aggregate measure from all approaches and retain only top k most likely to be relevant, prioritising those ranked highly by most approaches in combination. In our initial construction, we set k to a maximum of 50,000 publications based on pragmatic computational and financial constraints. This threshold provided a balance between comprehensive literature coverage and the practical limits of further automated and semi-manual labelling and processing. For this subset, we first classified as irrelevant any publications that lacked statistical tables presenting model coefficients, odds ratios, hazard ratios, or score computation heuristics. We then visually inspected up to 200 randomly selected articles from the remaining set, labelling each as relevant or irrelevant based on the detailed content, providing ground truth for supervised learning. Using a standard machine learning pipeline, we trained a simple meta-model with input features comprising distances between LLM embeddings of articles and queries, binary indicators from a dictionary of RE matches, and continuous probabilistic string match scores, using relevance labels as outputs. To improve data efficiency in relevance detection, we initialised the meta-model with parameters and processes adapted from Pharmatics' other commercial and scientific research projects. This approach allowed us to build upon our existing, high-quality screening method, enabling few-shot learning and reducing the need for extensive retraining of the relevance evaluation model from scratch. This enabled the model to learn an effective combination of evidence features for task-specific relevance assessment in HCM with minimal additional labelling.

Processing of Tabular Data

Further, the pipeline evaluated the actionability of articles by analysing the contents of statistical tables or published code. To detect and extract tables from images and PDF documents, we employed computer vision and image processing techniques, followed by a deep convolutional neural network (inspired by TableNet [2]) pre-trained on scientific articles to segment page regions and determine precise table boundaries. We then applied an established optical character recognition method

(Tesseract [3]) to cropped regions to extract table contents, converting them into a machine-readable format suitable for downstream analysis. For tables embedded in text or in structural documents such as HTML or XML, we extracted text directly using text parsers. After this multi-step extraction process, each table was linearised and represented as a dictionary of keys and matched values to provide a structured format for automated downstream processing. We then analysed these representations using a multi-stage classification cascade, which included an initial rule-based filter to discard obvious non-statistical tables. Then, supervised multi-class machine classifiers were used to categorise table contents into common table classes such as demographic summaries or model-specific outputs (for example, logistic regression or Cox proportional hazards model coefficients). At that stage, we used our internal dataset augmented with the manually labelled tables from the examined HCM papers. To perform quality control, we fine-tuned a domain-specific LLM using a multitude of previously annotated tables and linking linearised table contents to their corresponding captions and references in the surrounding texts; these LLMs were prompted to evaluate the internal consistency and quality of the outputs of supervised classifiers and extracted model coefficients and refine extraction if necessary. Empirically, we found that this combination of LLMs and cascade classifiers outperformed each approach used in isolation. It should be noted that the tabular data processing is being continuously extended to detect and classify challenging layouts, including complex multi-page tables, multiple sub-tables, split rows, columns, or cells, etc. Our initial library of evidence-based models used for preparing this deliverable includes approximately twenty typical model-based table layouts but is constantly growing.

Variable Harmonisation

One of our aims was to ensure consistency in variable names and measurement units across datasets, enabling multiple evidence-based functions constructed for different populations to share a standardised interface within the API wherever possible. To achieve this homogeneity, we developed a data harmonisation tool using LLM prompting to generate multiple synonymous natural language descriptions for each variable name from every HCM function constructed using the MedAI pipeline. For this, we provided the variable names as they appeared in the relevant tables of model coefficients, along with the wider context such as captions and related text referencing these tables. The resulting descriptions were converted into LLM embeddings and projected into a lower-dimensional feature space, with a contrastive objective [4] optimised to ensure that synonymous descriptions of the same variable were positioned closer in the feature space than unrelated descriptions. This model enabled the identification of similar variables across different functions that represent the same concept. We then used the resulting similarity metric as the input for an agglomerative clustering algorithm, grouping variables from distinct functions that correspond to the same underlying concept. Each cluster was assigned a unique internal canonical variable name, together with standardised metadata such as typical value ranges, physically plausible limits, and measurement units. Semi-manual curation was performed to address edge cases, such as the coding of binary and multinomial categories and the expansion of common clinical abbreviations within variable names. In practice, this approach proved effective for harmonisation of publication results, as the probability produced by the trained contrastive model was specifically calibrated for clinical variable descriptions and similarity assessment. This contrasts with relying on the common cosine (scalar product) similarity between raw embeddings from domain-independent language models, which empirically resulted in strongly right-skewed similarity distributions and frequently misclassified even unrelated clinical concepts as geometrically similar. In such cases, the cosine similarity values of general-purpose LLM embeddings tended to cluster near one, reducing discrimination between similar and dissimilar concepts. Furthermore, our current approach has the advantage of interpretability over geometric similarities: for example, a similarity probability of 0.95 can be directly understood as two variables having a 95% likelihood of being similar, an interpretation not possible with cosine similarity metrics. Despite encouraging initial results, our variable clustering and harmonisation tool is being continuously updated and fine-tuned as more data and edge cases become available.

Metadata Generation

For every evidence-based function extracted from the literature, we constructed detailed metadata to ensure standardisation, transparency, and interoperability across the work packages of SMASH-HCM. For each model, we included comprehensive metadata fields describing all input variables, specifying labels, variable names, data types, measurement units, descriptive text, and explicit constraints such as acceptable value ranges or categories. We documented the structure of the statistical model itself by recording the model type (e.g., logistic regression), listing all parameters with their names, values (such as odds ratios), and corresponding interpretations. The metadata also captured information about the study population from which the model was derived, including a textual description of the cohort and the country of origin. Each function was annotated with a unique label, a short descriptive summary of its purpose, and a distinct function name. Output definitions were included, detailing the type and meaning of each model output. This extensive internal documentation supports the interpretability, validation, use, and maintenance of each model; facilitates the harmonisation of variables; and enables the integration of evidence-based functions for downstream analyses.

Code Generation

To construct predictive models from extracted metadata for the MedAI-HCM library, we used the information contained within the structured table and variable annotations, converting them directly into *Python* code. For common statistical models such as linear regression, logistic regression, Poisson regression, and survival regression families (including Cox proportional hazards and Weibull models), this process was relatively straightforward. Each function's meta-data provided canonical variable names, variable types (such as integer or float), expected measurement units, and detailed descriptions, along with relevant input constraints like permissible value ranges or allowed categories. Using this information together with the harmonisation tool described above, we mapped non-canonical variables from the functions' metadata to standardised names, types, and measurement scales, ensuring consistency across functions, wherever possible. We then used the published model coefficients, odds ratios, or hazard ratios, typically specified in the statistical tables or appendices in publications, together with the model type and variables, to reconstruct the relevant predictive functions within the standard model families. For commonly used variables, automated measurement unit conversion libraries were used to allow input specifications using units that differ from the canonical forms, with the automated scaling of the underlying coefficients. By adhering to the canonical names given for each variable, the generated models could accept standardised input values and produce interpretable predictions, such as risk probabilities. For each model, the population meta-data was linked to the generated Python functions to support downstream calibration, model refinement, and relevance assessment for specific predictive scenarios, such as the risk prediction of clinical events or complications (for example, sudden cardiac death or heart failure in HCM or general populations), patient prognosis, and treatment response. Output specifications from the meta-data were used to define the model's return type and clarify the clinical interpretation of predictions, for example, expressing results as relative risk or the probability of a clinical event.

In contrast to standard statistical models commonly found in the medical literature, the generation of symptom score-based functions required a more involved process due to the greater heterogeneity of underlying published methods. Efficient and accurate score generation remains ongoing work and is achieved through a multi-stage process that includes automatic identification of the score type, parsing of the associated symptoms or questionnaire items, and extraction of the weights assigned to each item using separate logic for different common score types. This information is then used to reconstruct score computations as programmatic functions within the API. Overall, the number of model types that can be automatically classified and generated from extracted coefficients is continuously growing as the system is extended to recognise new statistical models and scoring methods from the literature. This approach has enabled the initial large-scale, automated synthesis and harmonisation of predictive models from published evidence, and will support applications across

data-driven analyses in WP6, clinical decision support development in WP7, and potentially a pilot study in WP8.

Model Calibration

The model generation approach described above had a major limitation: models reconstructed from the literature often lacked key parameters, most notably the bias terms or intercepts. Model calibration therefore needed to be performed to address this common issue in automatically extracted statistical models, such as logistic or linear regression, in which these terms were frequently omitted from published tables. Without these terms, the generated functions can only stratify patients by relative risk, rather than provide estimates of absolute risk or probability, which may be required for clinical decision-making and certain analytical tasks. To address this limitation, we systematically extracted additional information from the population (demographic) tables in the relevant publications. These tables typically contained the distribution of predictors within the studied cohorts and summary statistics for model outputs, such as the numbers of cases and controls, or distributions of continuous outcomes. Based on this information, we constructed a "median" patient, whose predictor values reflected the median or average for each variable within the ranges reported in the meta-data. We then applied this representative patient profile to each reconstructed model function and compared the uncalibrated model output with the actual cohort-level outcome, such as the observed event rate or median value of a continuous trait. The model's intercept or bias term was then recalibrated so that the prediction for the median patient matched the cohort's reported outcome. This approach projected model predictions onto the discovery population, automatically converting relative risk scores into absolute risks as reported in the literature, whenever possible.

Our approach to automated calibration is still evolving. One of the main challenges is accurately matching the representative patient to the model, especially when some relevant variables used as predictors in the models are missing from the population descriptions in the same papers; in such cases, we impute missing values using canonical distribution-based estimates. Another important challenge is the automatic calibration of survival regression models, which often requires reconstructing baseline risks from Kaplan-Meier curves. These curves are highly heterogeneous and require complex, multi-stage processing for automatic detection and information extraction. Although certain academic publications have claimed to achieve high-quality automated extraction of Kaplan-Meier curves from images [6], we were unable to reproduce those results in our datasets. To the best of our knowledge, this task remains largely unsolved. Our current approach for MedAI-HCM uses a cascade classifier that identifies and categorises common plot and curve features, such as presence of solid or dashed lines, multiple curves or subplots, error bars, shaded confidence intervals, colour coding, censoring marks, partial occlusions, etc., and applies dedicated machine learning procedures to digitise, smooth, and monotonically interpolate curves for a growing set of plot types. We are working on extending this methodology to accommodate less common plot formats. Another strategy we plan to explore later in the project involves recalibrating models using patient SMASH-HCM data, for those models whose input variables can be mapped to the SMASH-HCM data dictionaries.

Automated Generation of Example Use Cases

To test evidence-based functions from MedAI-HCM, we constructed clinical vignettes by sampling covariate values from the marginal distributions described in the population summary tables that accompanied each model's metadata. For each input variable, we sampled values based on the reported ranges, means, or medians, and rejected any sample that fell outside its physically plausible range, as defined by the metadata or canonical constraints (e.g. age below zero or extreme body mass index). For a subset of these samples, we systematically perturbed selected covariate values to verify that the function's risk estimates were consistent with established clinical expectations—for example, by incrementally raising the NYHA stage in HCM vignettes to confirm that the predicted risk increased accordingly. These cases were not intended to mirror real-world clinical narratives but rather to ensure that the generated functions produced sensible outputs: plausible risk estimates when given valid

inputs, and appropriate error messages or behaviour when inputs were omitted or outside the allowed range. In this way, the clinical vignette generation served as a practical validation tool to check the correctness, robustness, and interpretability of the automated model implementations. In the future, our current approach could be further enhanced by incorporating information about the joint distribution of covariates (when available) to generate more realistic and internally consistent vignettes, by applying generative models trained on individual-level SMASH-HCM datasets, and by integrating feedback from clinicians to ensure more informative use-case validation. This strategy will be prioritised for evidence-based functions that show promise for improving risk prediction and stratification in SMASH-HCM datasets and are considered as candidates for inclusion in the decision support system to be developed in WP7.

Automated Construction of Browsable Reference Manuals

To ensure that programmers and researchers could easily access and understand each automatically extracted evidence-based function from the MedAI-HCM library, we created a browsable programmer reference that includes models for HCM and its complications (which at present comprise atrial fibrillation, heart failure, and stroke). This was achieved by converting the detailed meta-data, summarised above and describing each function's properties, into a structured HTML reference. The meta-data and resulting HTML documentation capture variable names, types, measurement units, descriptive summaries, input constraints, and brief model descriptions. Using an automated pipeline, this information is converted into a tutorial-style layout similar to Python documentation, which includes expandable menus and hierarchical navigation. Each function is listed and linked in the sidebar's table of contents, enabling users to quickly find and review multiple functions. Upon selecting a specific function, users can access a detailed description, parameter list, and specifications for expected inputs and outputs. The process is fully automated: whenever a new function is extracted from newly published literature and meta-data is generated, the reference manual is instantly updated to incorporate the latest additions. The level of detail, such as outcomes, target population information, publication references, or calibration data, is configurable can be tailored according to developer requirements. This approach ensures that different user groups can be granted varying levels of access, depending on their affiliations, intended use, or subscription status. Additionally, the approach ensures that the browsable documentation is consistent, accurate, and up-to-date for all available functions, supporting ongoing research and development within SMASH-HCM and beyond.

2.1.2 Function Characterisation

To facilitate understanding and effective use of the evidence-based MedAI-HCM functions, we systematically characterised these functions by grouping them according to the types of clinical outcomes they predict. We achieved this by first extracting descriptions of each function and then generating semantic embeddings using LLMs, capturing the core meaning of each function. To mitigate issues with high-dimensional feature spaces, we reduced the dimensionality of the embeddings to balance the preservation of meaningful semantic structure and noise reduction. We then applied agglomerative clustering to the lower-dimensional embeddings, allowing us to automatically group functions with similar outcomes. To enhance accessibility, we used advanced language models to assign short, human-readable names to each group, summarising the main clinical concept addressed by the clustered functions. We visualised the function landscape in low-D, making the groupings transparent and interpretable. This approach enables research and clinical teams to quickly view the spread and focus of available predictive models, benchmark new models predicting certain outputs against existing ones, construct outcome-specific ensembles, and readily identify areas where further model development or literature review is needed due to sparse function coverage.

The main steps of the process are outlined below.

Function Discovery

The initial step focused on scanning a database of MedAI-HCM functions, each associated with a link to an individual Python file generated during the function creation process described in Section 2.1.1. Each file contained a function with a concise description summarising its purpose. (For example, a description could state "Prediction of the 5-year risk of sudden cardiac death in European patients with obstructive HCM".) These short descriptions provided the textual data necessary for downstream processing. We parsed, extracted, and aggregated the descriptions, linking each to its corresponding metadata to ensure reproducibility. Next, we transformed these descriptions into numerical embeddings using OpenAI's GPT family of LLMs to capture the semantic meaning of function descriptions. To facilitate efficient and meaningful similarity assessment, we reduced the dimensionality of the embeddings using standard dimensionality reduction techniques, treating the number of retained dimensions as a hyperparameter. This embedding construction translated the semantic information of the functions into machine-readable form while reducing noise, which is essential for robust distance-based clustering.

Function Clustering

Using the generated embeddings, we applied agglomerative hierarchical clustering, which iteratively merges individual data points, each initially forming its own cluster, based on a chosen distance metric, until a hierarchical tree-like structure is formed. In our initial implementation of function characterisation, we used cosine distance between reduced-dimensionality embeddings. (More advanced approaches based on the probabilistic outputs of contrastively trained self-supervised models will be considered in the future.) This method identifies natural groupings in the feature space, allowing us to categorise functions by outcome type and semantic similarity. To evaluate clustering performance and select hyperparameters, we used standard clustering metrics alongside a set of labelled *must-link* and *cannot-link* pairs, corresponding to function types that should or should not be grouped together. To further support interpretability, each cluster was automatically assigned a concise, human-readable label summarising its general clinical theme, automatically generated from the aggregated function descriptions within each cluster using OpenAI's GPT-4.1 API. This automated labelling facilitates interpretability of cluster contents and will allow users to navigate MedAI-HCM functions within each grouping.

Visualisation

To provide an intuitive understanding of the clustering results, we visualised the grouped embeddings in two dimensions using established data visualisation methods such as t-distributed Stochastic Neighbour Embedding (t-SNE) [5]. t-SNE is a commonly used visualisation technique that preserves local neighbourhood relationships by converting pairwise similarities between samples (in our case, function descriptions) into probabilities and minimising the divergence between high- and low-dimensional similarity distributions to maintain original cluster structure in the visualisation space.

(Mathematically, let $\boldsymbol{x}_i$ correspond to a high-dimensional text embedding and $\boldsymbol{y}_i$ – to its lower-dimensional projection. t-SNE defines the conditional probability that function j is a neighbouring function of $i \neq j$ using a Gaussian kernel between in $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the embedding space

$$p_{j|i} \propto \exp\left\{-|\mathbf{x}_i - \mathbf{x}_j|^2 / 2\sigma_i^2\right\}$$

with $p_{i|i} = 0$ and the joint $p_{ij} = p_{i|j}/2N + p_{j|i}/2N$, where N is the sample size. It further defines the similarity between lower-dimensional projections $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ as being proportional to a t-distribution with one degree of freedom if $i \neq j$ with zero mass at $i = j$

$$q_{ij} \propto \left\{(1 + |\mathbf{y}_i - \mathbf{y}_j|^2)^{-1}\right\} \text{ if } i \neq j, \text{ and } q_{ij} = 0 \text{ otherwise.}$$

To construct the projections, it numerically minimises the KL-divergence $KL(p||q)$ with respect to projection coordinates $\boldsymbol{y}_i$ for all samples.)

By visualising the clustered functions in the 2D space, we enabled users to explore relationships and structure within the evidence-based functions, making groupings potentially more interpretable.

To summarise, the initial function characterisation approach provides an intuitive and robust framework for organising MedAI-HCM functions by integrating both established and emerging AI/ML methodologies. To ensure reproducibility of the characterisations, from function discovery through to visualisation, we controlled random elements within key algorithms, including those in LLM-based embedding generation, clustering, and numerical optimisation. Future work will involve further experimentation with hyperparameter optimisation, automated description generation, incorporation of supervised constraints such as *must-link* and *cannot-link* function pairs, and the use of probabilistic metrics derived from contrastively trained self-supervised models for function descriptions [4]. This methodology is expected to facilitate the identification, orchestration, and reuse of evidence-based models within SMASH-HCM and potentially in broader applications.

2.2 Population Analysis

Due to substantial genetic and phenotypic heterogeneity, reliable risk stratification remains a OKsignificant challenge in HCM. To address this, we performed an initial prognostic risk assessment of HCM, with particular emphasis on the role of sarcomere gene variation and insights gained from multi-modal phenotypic characterisation using unsupervised clustering techniques. This analysis serves as a first step towards identifying heterogeneous HCM subpopulations that may benefit from targeted application, validation, and refinement of evidence-based predictive models developed from biomedical literature and implemented in the MedAI-HCM pipeline. Separate calibration and refinement of these models will be carried out for each identified subtype in Task 6.3 and Task 6.4, with evaluation of model performance within each subtype in Task 6.5.

HCM is primarily a genetically inherited cardiovascular disease characterised by unexplained left ventricular hypertrophy in the absence of secondary causes such as hypertension [7,8]. It is regarded as one of the most common monogenic heart disorders [9], with an estimated prevalence range from 0.16%-0.29% in the general adult population [10–13]. HCM displays a broad spectrum of clinical expressions, ranging from asymptomatic individuals [14] to those experiencing severe symptoms such as exertional dyspnoea, chest pain, syncope, fatigue, palpitations, sudden cardiac death or heart failure [15–17], particularly in young athletes [18]. The disease is primarily caused by mutations in at least eight genes encoding sarcomeric proteins, most notably MYBPC3 and MYH7 [19], even though considerable genetic and phenotypic heterogeneity is present [9,15,20]. These genetic mutations initiate a cascade of abnormalities including disorganised myocyte architecture, interstitial fibrosis, myocardial hypertrophy, diastolic dysfunction, and left ventricular outflow tract obstruction [21,22]. While hereditary profiling has further facilitated the early identification of at-risk individuals, including asymptomatic family members [23,24], advances in imaging techniques have enabled more timely clinical intervention and precise risk stratification. In particular, cardiac magnetic resonance (CMR) imaging has emerged as a valuable non-invasive tool for detecting and quantifying myocardial fibrosis, which present in approximately 66% of HCM patients [25–27], thereby enhancing diagnostic accuracy and informing clinical decision making [28].

Despite advances in imaging and genetics, HCM continues to demonstrate substantial functional and structural heterogeneity. This complicates risk stratification, prognostic assessment of disease progression, and treatment effectiveness. To address these challenges, we are developing data-driven stratification models that classify patients with HCM and at-risk relatives into discrete risk groups, utilising multi-modal clinical data. Our primary objective in patient characterisation is to determine whether the resulting phenotypic clusters are associated with different levels of clinical risk, particularly across three major genotypic subgroups. Furthermore, we aim to establish a methodological framework for detecting transitions between risk clusters over time by analysing temporal dynamics in the most recent clinical data, thereby providing insights into disease progression and the impact of therapeutic interventions.

For our initial patient characterisation, we analysed data from the Sarcomeric Human Cardiomyopathy Registry (SHaRe) database [29] maintained in cardiovascular centre of Florence, which aggregates comprehensive electrical health records, genetic, and phenotypic data from over 2,600 patients and family members affected by HCM. Together with genomic profile, the dataset contains a wide range of multi-modal health data collected during clinic visits, including demographic information, vital signs, imaging modalities, such as echocardiography, CMR, electrocardiography (ECG), cardiac catheterisation, medication usage, stress test results, ambulatory monitoring, and other lifestyle factors. In addition, familial cardiovascular history and comprehensive records of cardiac events, such as prior occurrences of atrial fibrillation and premature ventricular contractions were incorporated to support more robust phenotypic clustering. All extracted variables were presented in a structured, tabular format. Features derived from various modalities, such as CMR, ECG or echocardiography were pre-extracted as quantitative measurements, allowing seamless incorporation into the modelling pipeline. From the genetic perspective, the dataset includes 8 sarcomeric genes and 40 others, along with details of their classifications such as Pathogenic (P), Likely Pathogenic (LP), Likely Benign (LB), or Variants of Uncertain Significance (VUS). The classification of mutation status was conducted according to the following criteria: SARC(+) indicates the presence of at least one P/LP variant in any of eight sarcomere genes, SARC(U) refers to cases with no P/LP variants but with at least one VUS in any of the sarcomere genes; and SARC(-) indicates the absence of P/LP and VUS variant in any genes. To enable temporal analysis, clinical parameters were summarised using aggregated statistical features, with a focus on minimum and maximum values, as well as the first and most recent measurements. Subsequently, summarised data were standardised and transformed to ensure compatibility with the analysis, followed by dimensionality reduction using principal components analysis and the reconstruction error criterion to determine a suitable threshold [30,31]. Three distinct clustering algorithms: K-Means, DBSCAN, and Gaussian Mixture Models were applied separately to each subgroup defined by sarcomeric mutation status to identify underlying phenotypic patterns within each genetic category [32,33]. Optimal parameter selection for the clustering algorithms was assessed through an equally weighted evaluation framework, consisting of three normalised validity metrics: the Davies Bouldin Index, Silhouette Score, and Calinski-Harabasz Index [34].

2.3 Software Development

To simplify usability and standardisation for integration into research workflows, we developed the MedAI-HCM API using the JSON-RPC protocol, a well-established lightweight standard for remote procedure calls. To provide secure access to a growing library of evidence-based predictive models (as described in Section 2.1.1), we implemented a dedicated API endpoint that enables calls to all available functions through a uniform interface, encoding both input data and responses in the simple, human-readable JSON format. In contrast to more complex protocols like SOAP, which often require complex XML-based message definitions, or REST, which typically relies on conventional HTTP methods and may require extra design efforts for complex or structured operations, JSON-RPC provides a straightforward and uniform interface for invoking predictive functions. This choice enables direct language-agnostic integration with a wide range of modern programming environments, including Python, R, JavaScript, and others. The simplicity and flexibility of JSON-RPC also reduce the potential for integration errors and overhead, while still supporting strong typing, error handling, and batching of multiple calls in a single request. Also, unlike REST, JSON-RPC allows predictive functions to be accessed through a consistent, uniform interface, without having to design separate endpoints for each function or to fit structured clinical data into URL formats. This choice makes it easier and more intuitive for research teams to submit structured or batched inputs to predictive models, facilitating the seamless integration of evidence-based analytics into research pipelines and clinical decision-support software. This design is particularly suitable for MedAI-HCM, where a large and evolving library of prediction functions may need to be updated without frequent changes to API endpoints.

To demonstrate and test the use of the API, we developed a complete Python example, `run_example.py`, which was made available to the development team. This script reads example requests from JSON files, encodes user credentials for server authentication, and securely sends function calls to the API endpoint via HTTP POST requests. The responses are then printed in a human-readable form or further processed as needed. Configurable server settings, such as host address, port, endpoint path, and authentication details, are managed separately in a configuration file (`example_config.py`). This approach ensures that users can easily adapt the interface to different environments without altering the core logic.

Function calls are defined in standard human-readable JSON files, each corresponding to a specific evidence-based function available in the MedAI-HCM instance. These files indicate which function to call and specify the required input parameters, typically retrieved from real datasets or synthetic clinical vignettes, including both values and measurement units for predictor variables that serve as model inputs for one or more patients. For example, a function call might request the prediction of all-cause mortality risk in an HCM patient by passing values for age, sex, NYHA class, presence of atrial fibrillation, time to event, and other required inputs of the corresponding evidence-based survival regression function from the MedAI-HCM API instance. Running the API examples is straightforward: executing `python run_example.py -a` will automatically execute all available example calls, while `python run_example.py --json <function_example.json>` allows targeted testing of specific function calls and inputs. The API responds with predicted values or risk estimates, as defined by the called model.

All function arguments and acceptable value ranges are validated against the metadata, facilitating built-in assertions and automated tests to ensure that inputs and outputs remain within biologically feasible limits, which facilitates data quality and safety. This metadata accompanies each MedAI-HCM function, providing comprehensive definitions for all required inputs, explanatory labels, data types, units, value constraints, and clinically meaningful descriptions. The metadata also records the statistical model types, key parameters (such as hazard ratios or odds ratios), their interpretations, and information about the study population and country of origin. Output types and their meanings are defined to ensure interpretability of results, potentially facilitating usability and uptake of the system.

To summarise, the MedAI-HCM API development is built around transparent, secure, reproducible, and well-documented interfaces, supported by a constantly evolving codebase that expands as new research is published. By providing comprehensive metadata, code examples, and seamless integration pathways, this methodology may enable robust adoption among preclinical and clinical researchers and healthcare IT teams, enabling timely access to peer-reviewed, literature-based predictive models in HCM and related conditions. In the future, we plan to extend the functionality by allowing users to invoke selected or all functions that match a given data dictionary. This means the system will be able to automatically identify a subset of functions from the API that are relevant to a specified outcome, such as sudden cardiac death, and are compatible with the available input variables. Functions that do not match the specified condition or cannot run due to input incompatibility will be automatically excluded. This may reduce manual effort and minimise errors due to variable mismatches, which can help researchers to focus on outcome-specific predictions that are best suited to their available datasets.

3 Results

This section presents the initial results of complementary streams of work: evidence-based function characterisation from the MedAI-HCM API instance (Section 3.1) and population characterisation using the SHaRe dataset (Section 3.2). This is followed by a brief overview of the current version of the MedAI-HCM API, which covers evidence-based functions for HCM and commonly associated complications such as atrial fibrillation, heart failure, and stroke. As the project is still in its early stages, we anticipate that both the characterisations and the API will evolve and be enhanced as development progresses.

3.1 Initial Function Characterisation

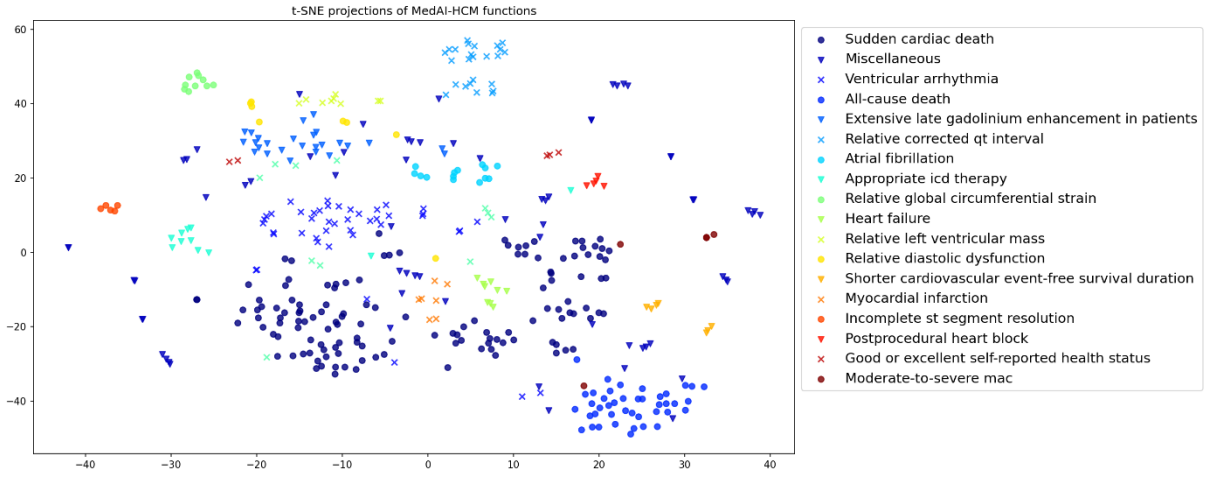

Figure 2 t-SNE visualisation of evidence-based functions from the initial instance of MedAI-HCM. Each point represents a single function, with colours and markers indicating clusters that were automatically detected according to common clinical themes based on the clustering of LLM-generated embeddings of function descriptions. Each cluster contains at least five functions. Smaller groups predicting disparate or less common outcomes were combined into cluster *Miscellaneous*. The plot illustrates common clinically relevant themes in HCM and related conditions, with the largest proportion of functions apparently focused on predicting sudden cardiac death. Functions predicting similar outcomes tend be grouped in close spatial proximity, while the diverse *Miscellaneous* cluster is more dispersed due to including a variety of distinct predictive outcomes.

Figure 2 illustrates the t-SNE projections of 446 functions extracted from approximately 20 distinct table layout types in the literature and incorporated into the current instance of the MedAI-HCM API. These functions include classification, regression, and survival regression models for a range of outcomes in patients diagnosed with HCM, those at risk of HCM, or at risk of common complications of HCM, such as atrial fibrillation, heart failure, and stroke. The models were synthesised from published literature following the process outlined in Section 2.1.1 and were categorised using the function characterisation method described in Section 2.1.2. No constraints were imposed on the data dictionaries, geographies, or interventions.

The process resulted in 18 clusters, each containing at least five functions, with the additional *Miscellaneous* cluster comprising functions that predict multiple disparate outcomes. The largest currently determined clusters were *Sudden Cardiac Death* (134), *Miscellaneous* (69), *Ventricular Arrhythmia* (44), *All-cause death* (42), *Extensive late gadolinium enhancement* (26), and *QTc prolongation* (25). Table 1 illustrates examples of cluster constituents for some of the clusters, showing 10 randomly selected functions from each cluster. Note that these cluster allocations and visualisations of the initial evidence-based functions are not perfect and have not yet undergone manual curation.

They are provided for illustrative purposes only, and both the clustering and visualisation methods may evolve as we continue to optimise our characterisation approach, as described in Section 2.1.2.

Table 1 Randomly chosen cluster constituents for several clusters of MedAI-HCM functions.

Cluster *Sudden Cardiac Death*
positive family history of sudden cardiac death and cardiovascular disease (SCD + CVD)
cardiovascular death or cardiac transplantation
sudden cardiac death
sudden cardiac death (SCD)
sudden cardiac death (SCD)
sudden cardiac death (SCD)
sudden cardiac death (SCD)
sudden cardiac death (SCD)
sudden cardiac death (SCD), ventricular tachycardia (VT), or ventricular fibrillation (VF)
ventricular arrhythmia (VA) and/or sudden cardiac death (SCD)
Cluster *Atrial Fibrillation*
atrial fibrillation
atrial Fibrillation
postoperative atrial fibrillation (POAF)
postoperative atrial fibrillation (POAF)
atrial fibrillation (AF) recurrence
atrial fibrillation (AF) recurrence
atrial fibrillation (AF) recurrence
freedom from atrial tachyarrhythmia, derived from 416 non-PAF ablations data at a London centre
Relative new onset atrial fibrillation (AF)
Relative new onset atrial fibrillation (AF)
Cluster *Appropriate ICD Therapy*
appropriate ICD therapy
incidence of appropriate ICD therapy administration
appropriate ICD/CRT-D therapy
ICD discharge
first appropriate ICD therapy
inappropriate shock due to myopotential interference (IAS) post S-ICD implantation
appropriate ICD shock
primary (PP) patient with an ICD versus being in the control group without an ICD
the appropriate implantable cardioverter defibrillator (ICD) therapy
the first occurrence of an appropriate shock after implantation of the ICD
Cluster *Miscellaneous*
in-hospital mortality
adverse events (AE)
hypertension in subjects who have never suffered an acute myocardial infarction (AMI) in the past
acute coronary syndrome (ACS)
prediction of male sex
abnormal MTWA
impaired apical contraction
flecainide Challenge Test (FCT) positivity among patients with non-type 1 Brugada ECG pattern
ventricular arrhythmia during hospitalisation
predicted right ventricular failure following left ventricular assist device (LVAD) implantation

At present, functions predicting composite endpoints are not assigned to a dedicated composite cluster; instead, they are grouped according to similarity metrics between their semantic LLM embeddings. For example, a function predicting "sudden cardiac death or implantable cardioverter-defibrillator" may be assigned to either the "sudden cardiac death" cluster or the "implantable cardioverter-defibrillator" cluster based on its pairwise distances to other cluster members, with its projection likely to appear near the boundary between these groups. This approach may be revised

later in the project, depending on end-user requirements for the research and development tool, which will be specified in Deliverable 7.6 ("SMASH-HCM Solution as research tool requirements").

3.2 Initial Population Characterisation

Based on the evaluation framework for patient characterisation described in Section 2.2, an agreement was reached across all three genetic subtypes, indicating that optimal clustering was achieved for two clusters (K=2). K-Means outperforms GMM and DBSCAN by achieving superior cluster separation and greater intra-cluster compactness. Variation in cluster membership was observed across the three SARC subsets. In SARC(+), cluster 1 and cluster 2 included 355 (50.5%) and 348 (49.5%) patients, respectively. In SARC(-), 336 patients (56.4%) were in cluster 1 and 260 patients (43.6%) in cluster 2. In subset SARC(U), 89 patients (51.4%) were in cluster 1 and 84 patients (48.6%) in cluster 2. The post-hoc analysis revealed exclusive stroke and abnormally elevated heart failure- and sudden cardiac death-related mortality in one of the clusters (Cluster 2) for all genotypic profiles, particularly among mutation-positive individuals, with 26 reported heart failure cases and 15 sudden cardiac death cases, compared to only one heart failure case and three sudden cardiac death cases in the other cluster (Cluster 1) for that subtype.

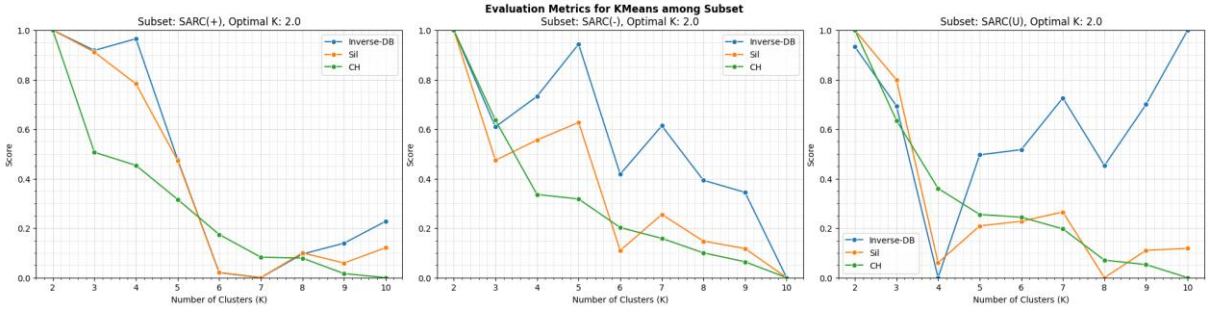

Figure 3: Clustering performance metrics for K-Means across the subsets. *Inverse-DB: Inverse of normalised Davies-Bouldin Index, Sil: normalised Silhouette score, CH: normalised Calinski-Harabasz Index.* Higher values indicate better clustering performance for all three metrics. The plot demonstrates consensus for optimal clustering at K=2.

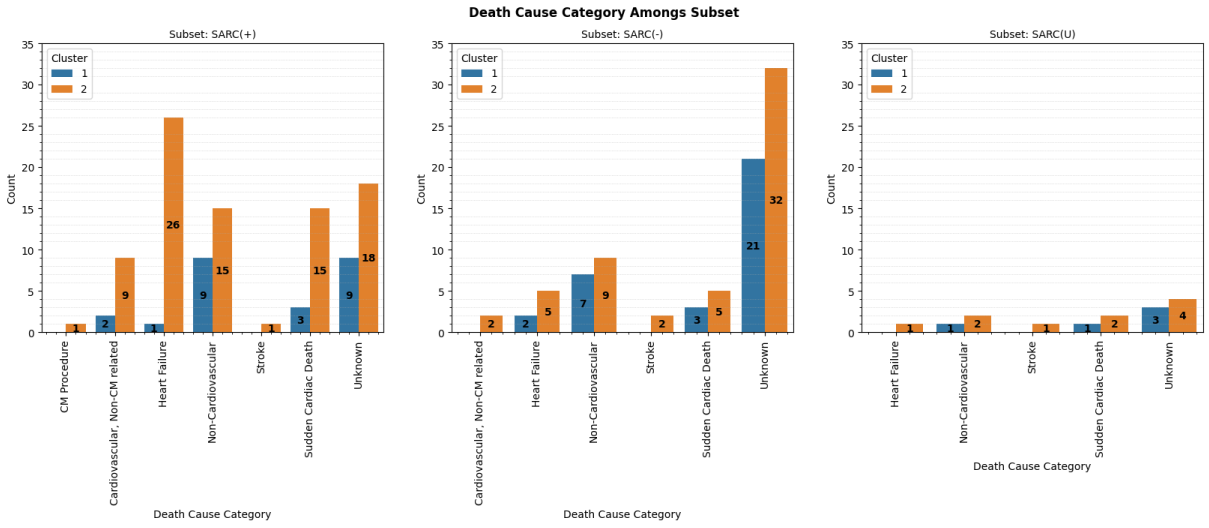

Figure 4: Distribution of death cause categories among subjects within each cluster for each genetic subtype. The results indicate consistent differences in death causes across the subtypes.

A comparative analysis of cumulative all-cause mortality conducted within the observed follow-up period, with medians ranging from 14.7 to 18.3 years across the three genetic groups revealed

consistently higher mortality in Cluster 2 relative to Cluster 1. Specifically, in the SARC(+), mortality was 24.4% in Cluster 2 and 6.8% in Cluster 1; in the SARC(-) group, 21.2% and 9.8%; and in the SARC(U) group, 11.9% and 5.6%, respectively. Notably, when restricting the study to cardiomyopathy-related deaths by excluding unknown causes, non-cardiovascular, and cardiovascular but non-cardiomyopathy causes, the disparity between the two clusters became even more pronounced in all three sacomeric groups. Indeed, in the SARC(+) cohort, cardiomyopathy-related mortality is 11.75 times higher in Cluster 2 than in Cluster 1, and 3.4 to 4.3 times greater than in other two categorisations.

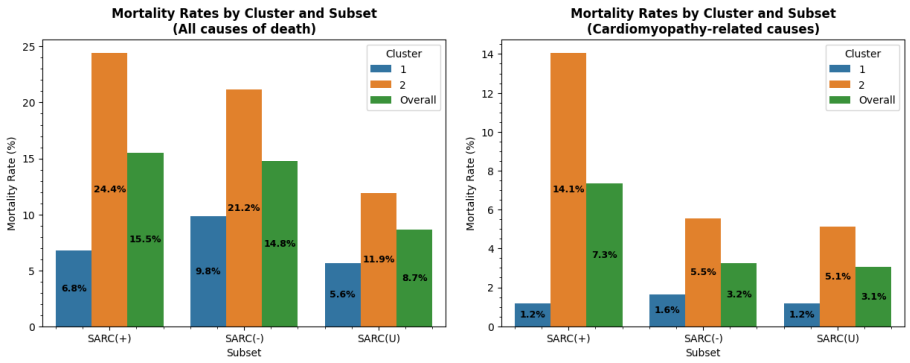

Figure 5: Comparison of all-cause *(left)* and cardiomyopathy-related *(right)* mortality across patient clusters and sarcomere categories: 1: Mortality rate for Cluster 1 (blue); 2: Mortality rate for Cluster 2 (amber); Overall: Mortality rate for the entire genetic subtype (green).

In addition, consensus across mutation groups indicates Cluster 2 is associated with an earlier reduction in survival probability, suggesting more aggressive disease progression during earlier life stages. Cluster 2 is associated with shorter lifespans, with forecasted values ranging from 75.2 to 80.2 years. This range is notably lower than the average predicted lifespans in other groups, which are 81.3 years for SARC(+), 84.0 for SARC(-), and 81.3 for SARC(U). Individuals in Cluster 1 who carry P/LP variants in at least one sarcomere gene exhibit the highest observed age at death, with a maximum recorded age of 90 years.

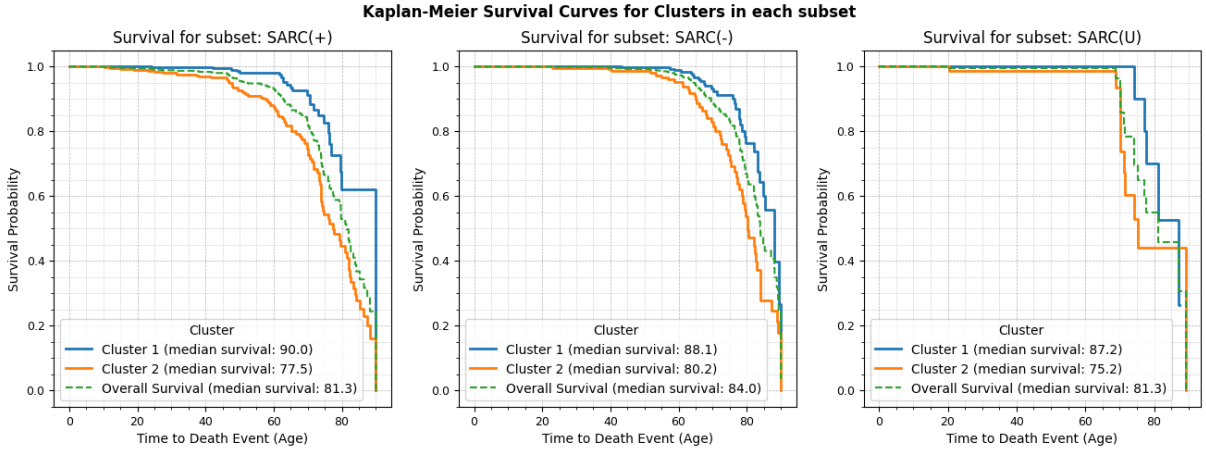

Figure 6: Kaplan-Meier Survival Curves for Clusters in each subset. 1: Survival curve for Cluster 1 (blue), 2: Survival curve for Cluster 2 (amber), Overall: Survival curve for the entire subset, regardless of cluster (green). The result suggests substantial differences in survival times between the clusters across the considered genetic subtypes.

The interpretation of clustering outcomes was guided by an analysis of the factor loadings associated with the first three principal components, which together captured most of the variance in our input data. Variables with high absolute coefficients on these components were found to substantially influence cluster separation, suggesting that the observed categorisations are primarily driven by differences in those features. While echocardiography and CMR data can be subject to rotational

variability, we do not expect this to affect the interpretability of PCA loadings in our study, as all images were acquired under standardised protocols that constrain orientation. As a result, we observed a dominant contribution of CMR and stress test parameters in the absolute coefficients of the top three components within the SARC(+) and SARC(U), while SARC(-) acknowledged greater contributions from echocardiographic and CMR variables. Among these, indices related to blood pressure, end-systolic volumes, end-diastolic volumes, and ejection fraction for both the left and right ventricles, demonstrated the strongest influence.

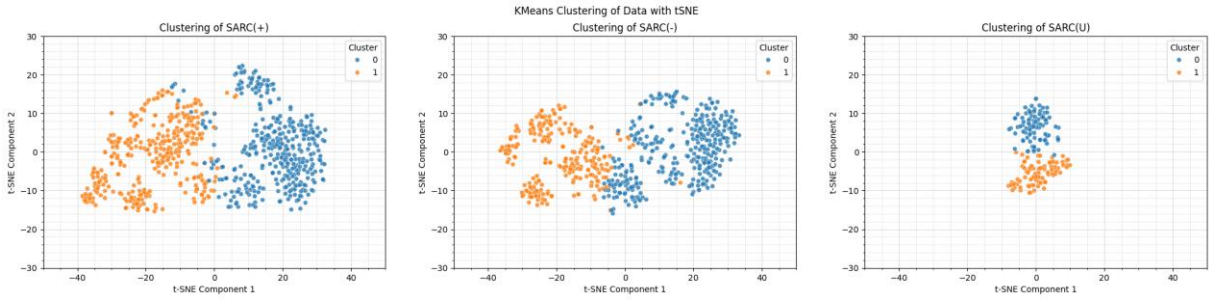

Figure 7: Visualisation of clusters in 2D using t-distributed Stochastic Neighbour Embedding (t-SNE). The plot demonstrates well-separated clusters across all considered genetic subtypes.

Our current result is the initial step towards HCM prognostic risk assessment, demonstrating that individuals allocated to one of the clusters exhibit significantly higher mortality rates, shorter estimated life expectancy, and noticeable elevated risks of adverse outcomes, such as sudden cardiac death, heart failure, or stroke, especially among individuals carrying P/LP sarcomeric variants. Furthermore, the initial findings also indicate that the utility of sarcomeric mutation status as a standalone stratification factor may be limited due to the consistency of risk profiles among individuals in the same cluster, regardless of the genetic subtype, thereby confirming the rationale for constructing a multimodal approach in HCM risk stratification. By assigning new samples to the cluster whose centroid is closest in Euclidean distance, the model enables the classification of unseen data points based on previously learned group structures, thereby facilitating the tracking of cluster transitions over time by incorporating recent medical records of individual patients. The proposed clustering approach not only reveals underlying risk features among the patients included in this study for model refinement and calibration but also lays the foundation for real time classification of other cases by proximity to learned centroids, paving the way for its integration into clinical decision support systems.

3.3 MedAI-HCM API

This section provides an overview of the MedAI-HCM API, which integrates evidence-based predictive functions characterised in Section 3.1 and designed for distinct clinical endpoints. The API's modular structure allows these functions, each associated with a particular clinical outcome cluster, to be validated and refined for specific patient subgroups identified in the population characterisation of Section 3.2. Each model included in the API can be adapted according to the demographic and clinical features of its relevant population clusters, which may support a stratified approach to risk prediction.

The currently constructed and partially calibrated instance of the MedAI-HCM API contains a total of 446 functions predicting various outcomes in patients diagnosed with HCM, as well as in individuals without a confirmed diagnosis of HCM but at risk of HCM or related conditions, such as atrial fibrillation, heart failure, and stroke. These models include only multivariate functions (specifically, those derived from the joint, or so-called *adjusted* analyses, instead of those reconstructed from univariate association studies as is common in the development of polygenic risk score models). Rather than presenting several hundred functions in this deliverable report, we provide a selection of screenshots from the automatically generated browsable reference manual. We then include the technical specification template, which was developed to provide extended information for all

functions developed in the SMASH-HCM project, including fields absent from the MedAI-HCM metadata and tailored to the needs of project partners. Finally, we refer to extended technical specifications for several selected functions.

3.3.1 Programmer Reference Manual

As described in Section 2.1.1, we have developed software to automatically generate browsable programmer reference manuals from meta-data. The MedAI-HCM reference is available to project partners. This section includes several snapshots of drop-down menus for randomly selected functions.

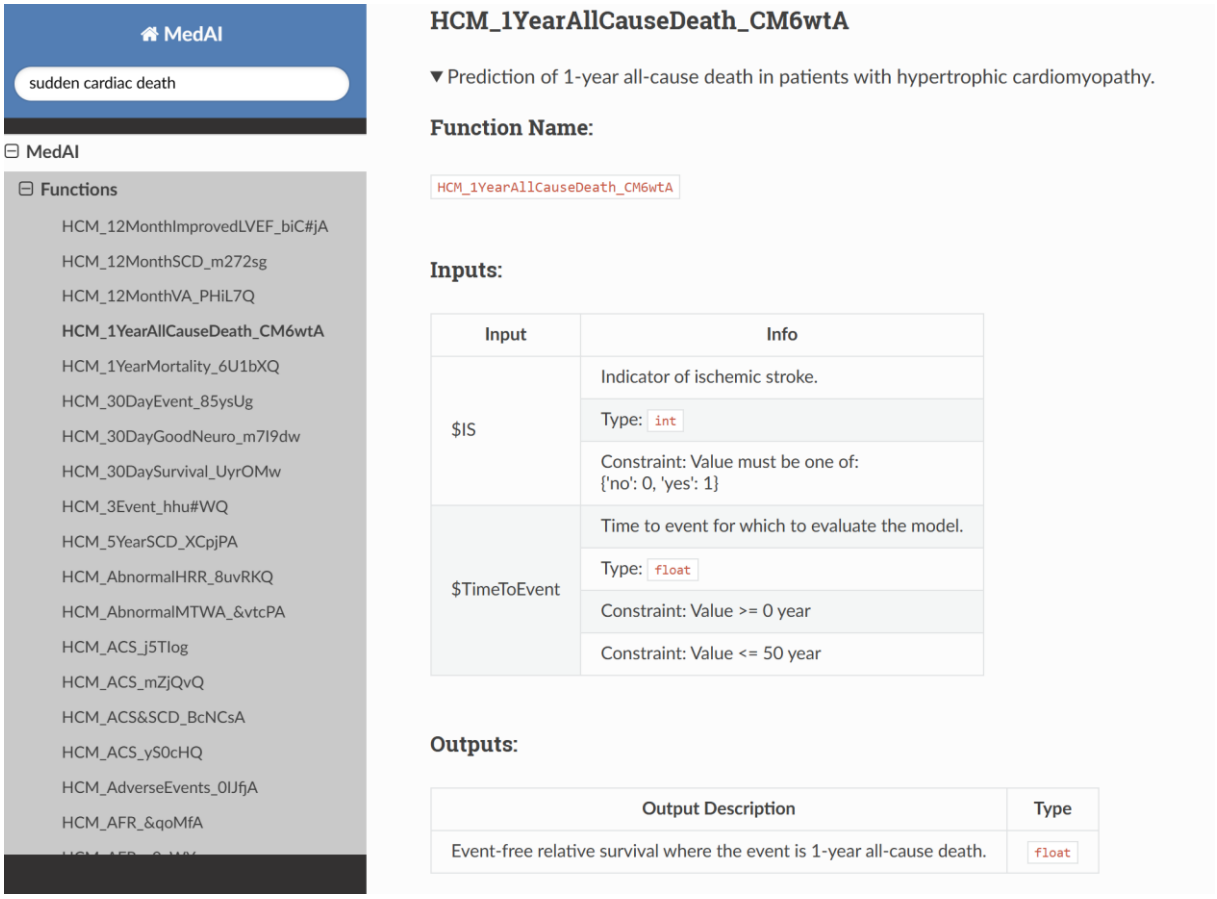

Figure 8 Snapshot of the programmer reference manual for a survival function based on IS history.

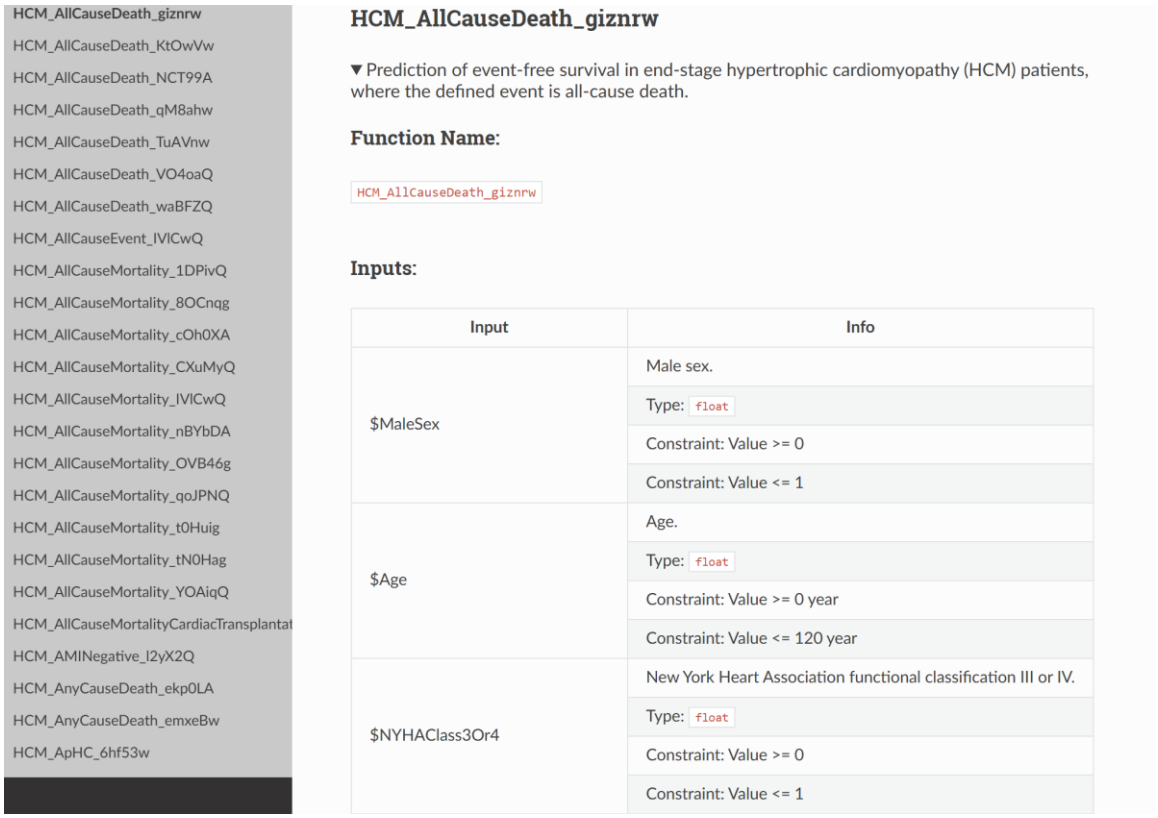

Figure 9 Snapshot of the programmer reference manual for an all-cause death prediction function.

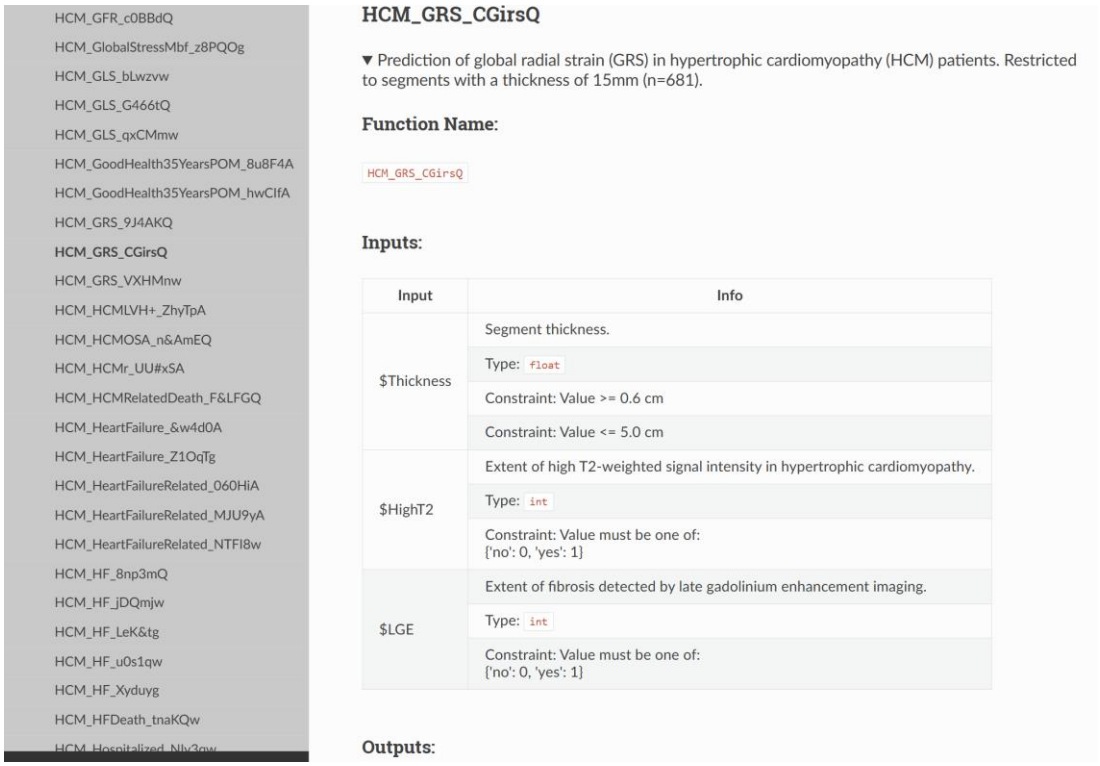

Figure 10 Snapshot of the programmer reference manual for a GRS prediction function.

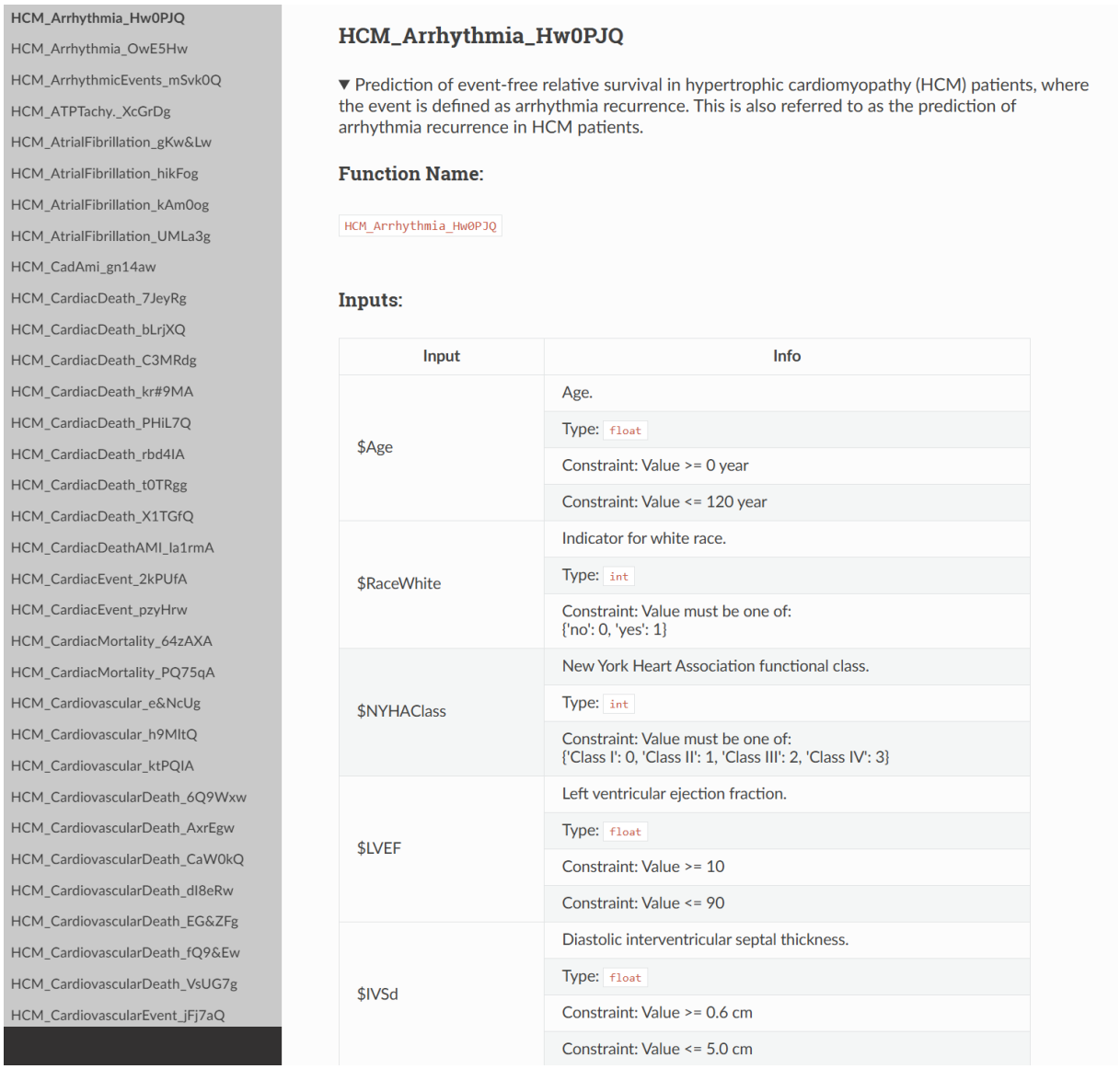

Figure 11 Snapshot of the programmer reference manual for a survival regression model predicting time to arrhythmia recurrence.

3.3.2 Technical Specification Template

In Deliverables D7.3 and D7.4, we developed a Technical Specification Template and the initial API, incorporating models from multiple project partners to ensure that all models—both the evidence-based models constructed in this Deliverable 6.2 and those developed across various tasks and work packages within SMASH-HCM—adhere to a consistent specification. The purpose of this template is to facilitate the use and integration of multiple models by different groups within the consortium. Table 1 includes this template for completeness.

Table 2 Updated Technical Specification Template for the SMASH-HCM API.

Field	Explanation
Name of a function, model, or data collection protocol	The unique name or title of the function, model, or data collection protocol. This should be descriptive enough to convey the purpose or role of the function, model, or data collection.

Type of a function, model, or data collection protocol	The category or nature of the function, model, or data collection protocol, such as whether it is an AI/ML regression model, a classification model, a survival regression model, a numerical model based on FEM modelling, ODEs; for data collection protocols, is it a cross-sectional or longitudinal study, interventional or observational, etc.
Short description of a function, model, or data collection protocol	A concise summary of the function, model, or data collection protocol, including its main purpose or intended use. This description may be one or two sentences long and will appear in the header of the models and functions in the SMASH-HCM model browser.
Inputs of a function or model	Details about what the function/model needs as input, including: • **Name:** The identifiers for each input variable. • **Type:** The data type of each input (e.g., integer, float, string, dictionary, vector, matrix, tensor, etc.), or format of input files if applicable. • **Unit:** The measurement unit of each input (e.g., ms, µmol/L, mg/dL, kg/m^2, ml/min/1.73m^2, etc.) • **Description:** A brief explanation of what each input variable represents. • **Range:** The range of plausible values each input can take. In case of categorical variables, define all categories the input might have and their meaning. For data collection protocols, specify variables or parameters of experiments that can be used for indexing the results (such as cell lines, temperature, pH, etc.)
Outputs of functions, models, or data collection protocols	Details about what the function, model, or data collection produces as outputs, including: • **Name:** The identifier for the output variable. • **Type:** The data type of the output (e.g., integer, float, string, dictionary, vector, matrix, tensor, etc.), or format of output files if applicable. • **Unit:** The measurement unit of the output (e.g., ms, µmol/L, mg/dL, kg/m^2, ml/min/1.73m^2, etc.) • **Description:** A brief explanation of what the output represents. • **Range:** The range of plausible values an output can take. In case of categorical variables, define all categories the output might have and their meaning.
An extended description of a function, model, or data collection protocol.	A more detailed explanation of the function, model, or data collection protocol, including its theoretical background, underlying algorithms, assumptions, relevant mathematical formulations, or clinical evidence. This description may be similar to the method paragraph in the abstract of a scientific publication and include multiple sentences. It will appear under the function description tag in the API reference if an investigator decides to examine a specific model in more detail.
Function, model, or protocol version (if applicable)	Version number of the function, model, or protocol.
References to data sources (if applicable)	Links to publications or other resources describing datasets or data collection protocols used in the development or evaluation of the function or model, including version numbers or dates if applicable, and data credibility evaluations (assessment of measurement errors,

	data provenance, quality, metrological properties, instrument or software certifications) if applicable.
References to labels or annotations (if applicable)	Information on the labels or definitions of the output variables or annotations that the model is designed to estimate, including references to sources, protocols, or definitions if applicable.
References to computational methods (if applicable)	Links to publications or other online resources describing the methods, algorithms, architectures, learning or inference techniques used in the development of the function/model, including academic papers, documentation, or other resources.
Software (if applicable)	Details about the software environment required to run the function/ or model or carry out data collection: • **Programming language:** The programming languages used (e.g., Python, R, C++). • **Required packages/Libraries:** The packages or libraries needed, including specific versions (e.g., NumPy 1.19.2, TensorFlow 2.3.0). • **Configurations:** Configuration files or settings required for running the function/model, if applicable. • **Environment variables:** Any necessary environment variables that need to be set for the function/model to run correctly, if applicable. • **Licensing agreements:** Any licenses or terms of use associated with the software or model. • **Installation instructions:** Any relevant instructions.
System and computing requirements (if applicable)	• **Operating system:** Windows, Linux, Mac OS, others. • **Hardware:** minimum hardware required (e.g., CPUs, GPUs, RAM, etc.) • **Batch mode:** if the simulation can or needs to be launched in batch mode • **Execution:** parallel execution, single-/multi-threading • **Running times:** expected duration of simulations
Intended use	Intended use of the model or research outputs, such as preclinical or clinical, including optional details.
Detailed clinical use (optional)	Clarification of the clinical question that may be addressed by using the model, including the population being studied, the predictors or interventions being analysed, the comparison or reference group (if applicable), and the outcomes being predicted or evaluated. • **Population:** The specific demographic or clinical group for which the model is intended. • **Intervention:** The specific treatment, device, or action that the model is designed to predict or inform. • **Comparison**: Any comparative metrics, control groups, or baseline metrics used in the model, if relevant. • **Outcome / Clinical Endpoint:** The primary results or endpoints the model is designed to predict or improve. • **Inclusion:** detailed patient characteristics required for eligibility to use the function, model, or resulting intervention. • **Exclusion:** specific characteristics that disqualify patients from using the function, model, or resulting intervention.

Methodological limitations	Any known constraints or weaknesses of the function, model, or data collection, including conditions under which it may not perform well, or specific scenarios where its predictions may be inaccurate.
Example of use (optional)	An example of a function or model call, or a brief description of a specific instance or scenario in which the research outputs of the function, model, or data collection could be applied. This might help users understand practical applications and the context of usage and help us describe research outputs appropriately for target audiences.
Additional Information	Any other relevant details that might explain the use of the function, model, or protocol or affect its performance or outputs. This may include information about lab conditions; local or cloud-based deployment environments; additional recommended technical or hardware specifications; best practices for using the function, model, or research outputs; considerations for handling missing or noisy data; interoperability requirements; compliance with standards or regulations; or any legal restrictions, if applicable.
Reference to evaluation	Links to publications or other online resources describing the quality or performance of the function or model in any scientific channel, including descriptions of the datasets used for evaluations, if applicable. (Note: initially, this field is likely to be empty for most models.)
Support team contact details	Contact details of the support team or principal investigators.

It should be noted that the template shown here and developed in Deliverable 7.4 is evolving as the project progresses, and future API instances of the evidence-based models will be documented according to the most recent technical specification requirements.

Following model construction, we converted the MedAI-HCM function meta-data into technical specifications, in accordance with the initial technical specification developed in Deliverables 7.3 and 7.4. For this conversion, we initially considered supervised fine-tuning of an LLM from the OpenAI GPT-4.1 mini family but eventually chose few-shot prompting with GPT-4o due to its greater efficiency in transforming the MedAI-HCM metadata into the SMASH-HCM technical specification template format for the current set of functions. Although this conversion was performed for the full current set of HCM functions, only a subset shown below was formatted in the *docx* format and included in this report due to space and time limitations for illustration purpose. Function calls themselves were tested either locally or on a virtual private server accessible to project participants for internal development and testing. Each function was verified to return the expected outputs or to fail gracefully when provided with invalid inputs. This section should be regarded as a living document, with additional functions to be incorporated into the technical descriptions as the project progresses. The programmer reference manual containing function descriptions is available to project participants through our internal Teams folder. Technical specifications of several typical functions are illustrated in Appendix B: Technical Specification of Selected Functions.

4 Conclusions and Future Work

To summarise, Deliverable D6.2 describes the initial development and implementation of evidence-based predictive models for HCM and its major complications, MedAI-HCM, within the SMASH-HCM Project. This evolving API, built on advanced artificial intelligence and machine learning methodologies, information extraction, and data harmonisation, lays the foundation for robust risk prediction, function and population characterisation, refinement and recalibration, and integration with research and development processes and clinical decision support systems. As the initial set of evidence-based models continues to be refined and validated using SMASH-HCM datasets, the evolving library will provide an increasingly accurate and up-to-date resource to support both research and clinical applications within and beyond our project.

As we are still in the early stages, the initial API has focused on evidence-based models that could be extracted from the literature using the current version of the general-purpose evidence-mining pipeline, Pharmatics' MedAI. Models and research results requiring more complex, multi-stage extraction or HCM-specific curation will be incorporated into the live API as the pipeline evolves to accommodate these challenging cases. Future API extensions will support increasingly complex outputs, including results presented in heterogeneous layouts that require distinct ad hoc table classification, information extraction, metadata generation, and/or model refinement algorithms. We are also actively expanding the pipeline to support a wider range of model types, with particular emphasis on the integration of highly heterogeneous symptom scores and clinical decision trees. Additionally, we are continuously enhancing the robustness of variable harmonisation and automated calibration steps, especially for complex models such as the various types of survival regression, which are critical for predicting and preventing clinical events like sudden cardiac death in HCM. Further ongoing enhancements may include refining calibration techniques, developing specialised machine learning methods for extracting baseline survival functions from heterogeneous graphical representations such as Kaplan-Meier curves reported in publications, incorporating joint covariate distributions for more realistic clinical vignettes, and integrating human feedback from HCM specialists to maximise the clinical applicability and relevance of our tools.

In near future, as the MedAI-HCM library of extracted evidence-based models continues to expand, it will be essential to identify and prioritise a subset of these initial functions for rigorous validation and refinement in individual-level SMASH-HCM datasets. This will ensure that the most promising models are validated and calibrated for our targeted populations, strengthening predictive data-driven modelling in WP6, supporting clinical decision tool development in WP7, and facilitating clinical and industrial uptake in WP8. We envision that MedAI-HCM will enhance research capabilities within and outside the consortium and may support more effective preclinical and clinical research and decision-making for HCM, potentially extending to other cardiovascular conditions.

References

[1] Pharmatics, www.pharmaticsltd.com/MedAI/Pharmatics-MedAI-leaflet-2025.pdf, 2025.

[2] Paliwal SS, Vishwanath D, Rahul R, Sharma M. and Vig L. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 international conference on document analysis and recognition (ICDAR)*, IEEE 2019: pp. 128-133.

[3] Unnikrishnan R and Smith R. Combined Orientation and Script Detection using the Tesseract OCR Engine, 2009. In MOCR '09: Proceedings of the International Workshop on Multilingual OCR, 2009: 1-7, https://tesseract-ocr.github.io/tessapi/4.0.0/.

[4] Gui J, Chen T, Zhang J, Cao Q, Sun Z, Luo H, and Tao D. A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence: 2024.

[5] van der Maaten, L.J.P.; Hinton, G.E. Visualizing Data Using t-SNE. Journal of Machine Learning Research. 9: 2579–2605, 2008.

[6] Zhang JZ, Rios JD, Pechlivanoglou T. et al. SurvdigitizeR: an algorithm for automated survival curve digitization. BMC Med Res Methodol 2024; 24, 147. https://doi.org/10.1186/s12874-024-02273-8.

[7] Arbelo E, Protonotarios A, Gimeno JR, Arbustini E, Barriales-Villa R, Basso C, et al. 2023 ESC Guidelines for the management of cardiomyopathies: Developed by the task force on the management of cardiomyopathies of the European Society of Cardiology (ESC). Eur Heart J 2023;44:3503–626. https://doi.org/10.1093/eurheartj/ehad194.

[8] Ommen SR, Ho CY, Asif IM, Balaji S, Burke MA, Day SM, et al. 2024 AHA/ACC/AMSSM/HRS/PACES/SCMR Guideline for the Management of Hypertrophic Cardiomyopathy: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. Circulation 2024;149:e1239–311. https://doi.org/10.1161/CIR.0000000000001250.

[9] Marian AJ, Braunwald E. Hypertrophic cardiomyopathy: Genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. Circ Res 2017;121:749–70. https://doi.org/10.1161/CIRCRESAHA.117.311059.

[10] Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT, Bild DE. Prevalence of hypertrophic cardiomyopathy in a general population of young adults: Echocardiographic analysis of 4111 subjects in the CARDIA study. Circulation 1995;92:785–9. https://doi.org/10.1161/01.CIR.92.4.785.

[11] Maron BJ, Mathenge R, Casey SA, Poliac LC, Longe TF. Clinical profile of hypertrophic cardiomyopathy identified de novo in rural communities. J Am Coll Cardiol 1999;33:1590–5. https://doi.org/10.1016/S0735-1097(99)00039-X.

[12] Hada Y, Sakamoto T, Amano K, Yamaguchi T, Takenaka K, Takahashi H, et al. Prevalence of hypertrophic cardiomyopathy in a population of adult japanese workers as detected by echocardiographic screening. Am J Cardiol 1987;59:183–4. https://doi.org/10.1016/S0002-9149(87)80107-8.

[13] Zou Y, Song L, Wang Z, Ma A, Liu T, Gu H, et al. Prevalence of idiopathic hypertrophic cardiomyopathy in China: A population-based echocardiographic analysis of 8080 adults. American Journal of Medicine 2004;116:14–8. https://doi.org/10.1016/j.amjmed.2003.05.009.

[14] Tisha Suboc. Hypertrophic Cardiomyopathy - Cardiovascular Disorders - MSD Manual Professional Edition. MD, Rush University 2024. https://www.msdmanuals.com/professional/cardiovascular-disorders/cardiomyopathies/hypertrophic-cardiomyopathy (accessed May 13, 2025).

[15] Basit H, Alahmadi MH, Rout P, Sharma S. Hypertrophic Cardiomyopathy. StatPearls Treasure Island (FL) 2024;Publishing, 2025.

[16] Van Driel B, Nijenkamp L, Huurman R, Michels M, Van Der Velden J. Sex differences in hypertrophic cardiomyopathy: New insights. Curr Opin Cardiol 2019;34:254–9. https://doi.org/10.1097/HCO.0000000000000612.

[17] Kraft T, Montag J. Altered force generation and cell-to-cell contractile imbalance in hypertrophic cardiomyopathy. Pflugers Arch 2019;471:719–33. https://doi.org/10.1007/s00424-019-02260-9.

[18] Maron BJ, Haas TS, Ahluwalia A, Murphy CJ, Garberich RF. Demographics and Epidemiology of Sudden Deaths in Young Competitive Athletes: From the United States National Registry. American Journal of Medicine 2016;129:1170–7. https://doi.org/10.1016/j.amjmed.2016.02.031.

[19] Popa-Fotea NM, Micheu MM, Bataila V, Scafa-Udriste A, Dorobantu L, Scarlatescu AI, et al. Exploring the continuum of hypertrophic cardiomyopathy—from DNA to clinical expression. Medicina (Lithuania) 2019;55. https://doi.org/10.3390/medicina55060299.

[20] Doeblin P, Gebker R, Pieske B, Kelle S. Late onset apical hypertrophic cardiomyopathy: a case report n.d. https://doi.org/10.1093/ehjcr/ytaa493.

[21] Mattsson G, Magnusson P, Mattsson G, Magnusson P. Cardiomyopathy - Disease of the Heart Muscle. IntechOpen; 2021. https://doi.org/10.5772/INTECHOPEN.91489.

[22] Yiu KH, Atsma DE, Delgado V, Ng ACT, Witkowski TG, Ewe SH, et al. Myocardial structural alteration and systolic dysfunction in preclinical hypertrophic cardiomyopathy mutation carriers. PLoS One 2012;7. https://doi.org/10.1371/journal.pone.0036115.

[23] Maron BJ, Seidman JG, Seidman CE. Proposal for contemporary screening strategies in families with hypertrophic cardiomyopathy. J Am Coll Cardiol 2004;44:2125–32. https://doi.org/10.1016/j.jacc.2004.08.052.

[24] Gersh BJ, Maron BJ, Bonow RO, Dearani JA, Fifer MA, Link MS, et al. 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: A report of the American College of cardiology foundation/American heart association task force on practice guidelines. Circulation 2011;124. https://doi.org/10.1161/CIR.0b013e318223e2bd.

[25] O'Hanlon R, Grasso A, Roughton M, Moon JC, Clark S, Wage R, et al. Prognostic significance of myocardial fibrosis in hypertrophic cardiomyopathy. J Am Coll Cardiol 2010;56:867–74. https://doi.org/10.1016/j.jacc.2010.05.010.

[26] Bruder O, Wagner A, Jensen CJ, Schneider S, Ong P, Kispert EM, et al. Myocardial scar visualized by cardiovascular magnetic resonance imaging predicts major adverse events in patients with hypertrophic cardiomyopathy. J Am Coll Cardiol 2010;56:875–87. https://doi.org/10.1016/j.jacc.2010.05.007.

[27] Noureldin RA, Liu S, Nacif MS, Judge DP, Halushka MK, Abraham TP, et al. The diagnosis of hypertrophic cardiomyopathy by cardiovascular magnetic resonance. Journal of Cardiovascular Magnetic Resonance 2012;14:17. https://doi.org/10.1186/1532-429X-14-17.

[28] Liew AC, Vassiliou VS, Cooper R, Raphael CE. Hypertrophic Cardiomyopathy—Past, Present and Future. J Clin Med 2017;6:118. https://doi.org/10.3390/JCM6120118.

[29] SHaRe – SHaRe Project n.d. https://www.theshareregistry.org/ (accessed May 22, 2025).

[30] Joe Qin S, Dunia R. Determining the Number of Principal Components for Best Reconstruction. IFAC Proceedings Volumes 1998;31:357–62. https://doi.org/10.1016/S1474-6670(17)44953-6.

[31] Valle, Sergio, et al. “Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods.” Industrial & Engineering Chemistry Research, vol. 38, no. 11, 1999, pp. 4389–401, https://doi.org/10.1021/ie990110i.

[32] Amr T. Hands-on machine learning with scikit-learn and scientific Python toolkits: a practical guide to implementing supervised and unsupervised machine learning algorithms in Python. Packt Publishing Ltd.; 2020.

[33] Scitovski R, Sabo K, Martínez-Álvarez F, Ungar Š. Cluster Analysis and Applications. Springer International Publishing; 2021. https://doi.org/10.1007/978-3-030-74552-3.

[34] Sowan B, Hong TP, Al-Qerem A, Alauthman M, Matar N. Ensembling validation indices to estimate the optimal number of clusters. Applied Intelligence 2023;53:9933–57. https://doi.org/10.1007/S10489-022-03939-W.

Appendix A: Evidence-based Modelling in the Context of WP6

This deliverable D6.2 is based on the initial research and development of Task 6.1 – "Evidence-based Modelling". Elsewhere in Work Package 6, Task 6.2 will build on a subset of literature-based models from Task 6.1 to investigate associations between evidence-based models, routinely collected variables, and clinical endpoints, with the aim of developing interpretable referral models suitable for integration into routine clinical workflows. Patients identified as high-risk using these commonly available variables will undergo extended data collection to facilitate confirmation of prognosis and to inform treatment strategies, drawing on models from Tasks 6.3 and 6.4. By prioritising methods that are interpretable by design and rely exclusively on routinely collected data, Task 6.2 is positioned as a potential initial entry point for clinical adoption. Task 6.3 shifts focus to multimodal patient-specific biomarker discovery employing advanced machine learning analyses of clinical data sources, ECG signals, and cardiac imaging (CMR, echocardiography). This task will use long short-term memory (LSTM) neural networks, convolutional neural network (CNN) autoencoders, and transformer-based models. Novel and interpretable biomarkers will be identified using both hypothesis-driven approaches informed by MedAI and purely data-driven methodologies. Further, this approach will aim to extract explainable representations in inner layers, while enabling integration of prior knowledge by applying cascade architectures to data from both open-source and SMASH-HCM cohorts. Task 6.4 further integrates the models and features developed in earlier tasks, combining them through multilevel ensemble approaches to enable more granular patient stratification and disease endpoint prediction. This will be achieved by using semi-supervised and embedded clustering techniques for multi-organ and multi-level datasets. Finally, Task 6.5 addresses extended technical validation and iterative refinement of these models, adhering to the principles of trustworthiness, fairness, and robust generalisation, systematically identifying and addressing sources of bias and AI vulnerabilities across diverse datasets. Overall, work package 6 will define a modular and interpretable predictive modelling framework, in which the outputs from each component can strengthen and refine subsequent processes.

Appendix B: Technical Specification of Selected Functions

Table WP6.1. Information given by: Ansari, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_PositiveHypertrophicCardiomyopathy_E2dcUA.
Type of a function, model, or data collection protocol	Logistic regression.
Short description of a function, model, or data collection protocol	Prediction of positive genetic test in hypertrophic cardiomyopathy in a Brazilian tertiary centre cohort of index cases with HCM.
Inputs of a function or model	• **Name:** HCFH__2_. • **Type:** int. • **Unit:** None. • **Description:** Familial history of cardiac disease without HCM diagnosis. • **Range:** 0 (no), 1 (yes).
	• **Name:** AHF. • **Type:** float. • **Unit:** 1/min. • **Description:** Average heart frequency. • **Range:** 60-100 1/min.
	• **Name:** NSVT. • **Type:** int. • **Unit:** None. • **Description:** Presence of non-sustained ventricular tachycardia. • **Range:** 0 (no), 1 (yes).
	• **Name:** Age. • **Type:** float. • **Unit:** year. • **Description:** Age. • **Range:** 0-120 years.
Outputs of a function, model, or data collection protocols	• **Name:** probability_positive_test. • **Type:** float. • **Unit:** none. • **Description:** Probability of positive genetic test in hypertrophic cardiomyopathy. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of positive genetic test in hypertrophic cardiomyopathy in a Brazilian tertiary centre cohort of index cases with HCM. The function uses a Logistic Regression model.

Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:3995628.
References to labels or annotations (if applicable)	See PMC:3995628.
References to computational methods (if applicable)	The model employs logistic regression to predict the likelihood of a positive genetic test.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** Patients with HCM clinical diagnosis. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Probability of positive genetic test in hypertrophic cardiomyopathy.
Methodological limitations	The model is based on data from a specific patient cohort and may not generalise well to other populations. Prediction accuracy is dependent on variable quality and distribution.
Example of use (optional)	TBD.
Additional Information	The model may aid in identifying patients likely to have a positive genetic test, guiding further investigations and management. However, it is for research use only and not yet clinically validated.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.2. Information given by: Ansari, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_VA_wv35Gw.
Type of a function, model, or data collection protocol	Logistic regression.
Short description of a function, model, or data collection protocol	Prediction of ventricular arrhythmias (non-sustained and sustained ventricular tachycardia and aborted cardiac arrest) in hypertrophic cardiomyopathy (HCM) patients.
Inputs of a function or model	• **Name:** Age_years. • **Type:** float. • **Unit:** year. • **Description:** Age. • **Range:** 0-120.
	• **Name:** MWT_mm. • **Type:** float. • **Unit:** mm. • **Description:** Maximal wall thickness. • **Range:** 5-50.
	• **Name:** GLS. • **Type:** float. • **Unit:** None. • **Description:** Global longitudinal strain. • **Range:** -30 to 0.
	• **Name:** Mechanical_dispersion_per_10ms_increase. • **Type:** float. • **Unit:** None. • **Description:** Mechanical dispersion. • **Range:** 0-100.
	• **Name:** LGE_present_by_CMR_n85_yes_vs_no. • **Type:** int. • **Unit:** None. • **Description:** Indicator of late gadolinium enhancement by cardiac magnetic resonance imaging. • **Range:** 0 (no).
Outputs of a function, model, or data collection protocols	• **Name:** risk_of_VA. • **Type:** float. • **Unit:** none. • **Description:** Relative risk of ventricular arrhythmias (non-sustained and sustained ventricular tachycardia and aborted cardiac

	arrest) in hypertrophic cardiomyopathy (HCM) patients. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of ventricular arrhythmias in hypertrophic cardiomyopathy (HCM) patients. The function uses generalised linear binary classifier model based on certain patient characteristics.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC: 4871235.
References to labels or annotations (if applicable)	See PMC: 4871235
References to computational methods (if applicable)	The model employs multivariate logistic regression to estimate the risk of ventricular arrhythmias.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** Patients with HCM clinical diagnosis. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Probability of positive genetic test in hypertrophic cardiomyopathy.
Methodological limitations	The logistic regression model requires the assumption of a linear relationship between the independent variables and the log odds, which is an inherent limitation and can affect prediction accuracy.
Example of use (optional)	TBD.
Additional Information	Upon validation, this model may be used clinically to pinpoint HCM patients at an elevated risk of ventricular arrhythmias, potentially guiding preemptive care

	strategies. Presently, this model is designated for research purposes only.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.3. Information given by: Orchard, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_AllCauseDeath_giznrw.
Type of a function, model, or data collection protocol	Survival regression.
Short description of a function, model, or data collection protocol	Prediction of event-free relative survival in end-stage hypertrophic cardiomyopathy (HCM) patients, where the defined event is all-cause death.
Inputs of a function or model	• **Name:** Male_sex. • **Type:** int. • **Unit:** None. • **Description:** Male sex • **Range:** 0 (no), 1 (yes).
	• **Name:** Age_per_10_year_increase. • **Type:** float. • **Unit:** year. • **Description:** age. • **Range:** 18-100.
	• **Name:** NYHA_class_III_IV. • **Type:** int. • **Unit:** None. • **Description:** New York Heart Association functional classification III or IV. • **Range:** 0 (no), 1 (yes).
	• **Name:** Atrial_fibrillation. • **Type:** int. • **Unit:** None. • **Description:** Presence of atrial fibrillation. • **Range:** 0 (no), 1 (yes).
	• **Name:** LVEF_per_10_decrease.

	• **Type:** float. • **Unit:** None. • **Description:** Left ventricular ejection fraction. • **Range:** 25-80.
	• **Name:** LBBB. • **Type:** int. • **Unit:** None. • **Description:** Presence of left bundle branch block on electrocardiogram. • **Range:** 0 (no), 1 (yes).
	• **Name:** Abnormal_Q_wave. • **Type:** int. • **Unit:** None. • **Description:** Presence of an abnormal Q wave on electrocardiogram. • **Range:** 0 (no), 1 (yes).
	• **Name:** 2_risk_factors_for_sudden_death. • **Type:** int. • **Unit:** None. • **Description:** Presence of >= 2 risk factors for sudden death. • **Range:** 0 (no), 1 (yes).
	• **Name:** t. • **Type:** float. • **Unit:** Year. • **Description:** Time to event for which to evaluate the model. • **Range:** 0-5 years.
Outputs of a function, model, or data collection protocols	• **Name:** survival_probability. • **Type:** float. • **Unit:** none. • **Description:** Event-free survival where the event is all-cause death. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of event-free relative survival in end-stage hypertrophic cardiomyopathy (HCM) patients, where the defined event is all-cause death. This is complementary to predicting all-cause mortality in patients with end-stage HCM. The function uses Cox Regression model without time-dependent covariates.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:4733774.
References to labels or annotations (if applicable)	See PMC:4733774.

References to computational methods (if applicable)	The model employs multivariate survival regression to estimate the probability of event-free survival.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 99 end-stage HCM patients from a total of 1844 consecutive HCM patients from April 2002 to November 2013 at Fuwai Hospital, aged 52 ± 16 years old at entry. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Event-free relative survival where the event is all-cause death.
Methodological limitations	The model is based on data from a specific patient cohort and may not generalise well to other populations. The predictions are dependent on the accuracy of covariates. The model assumes proportional hazards over time, meaning the effect of covariates on the hazard is constant, which may not hold true in real-world settings and can lead to biased results if violated.
Example of use (optional)	See Appendix C.
Additional Information	After validation, the model can potentially be used in a clinical setting to predict all-cause mortality. This can help clinicians identify high-risk patients, enabling timely interventions and personalised management strategies. This can also support shared decision-making around therapies like ICD implantation or advanced monitoring. Additionally, the model may be used as a feature for predicting other HCM outcomes. In its current form, this unvalidated model can be used

	for research and development and is not intended for medical use.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.4. Information given by: Orchard, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_CVDeath_bSKXXA.
Type of a function, model, or data collection protocol	Survival regression.
Short description of a function, model, or data collection protocol	Prediction of event-free relative survival in end-stage hypertrophic cardiomyopathy (HCM) patients, where the defined event is cardiovascular death.
Inputs of a function or model	• **Name:** Male_sex. • **Type:** int. • **Unit:** None. • **Description:** Male sex • **Range:** 0 (no), 1 (yes).
	• **Name:** Age_per_10_year_increase. • **Type:** float. • **Unit:** year. • **Description:** age. • **Range:** 18-100.
	• **Name:** NYHA_class_III_IV. • **Type:** int. • **Unit:** None. • **Description:** New York Heart Association functional classification III or IV. • **Range:** 0 (no), 1 (yes).
	• **Name:** Atrial_fibrillation. • **Type:** int. • **Unit:** None. • **Description:** Presence of atrial fibrillation. • **Range:** 0 (no), 1 (yes).
	• **Name:** LVEF_per_10_decrease. • **Type:** float. • **Unit:** None. • **Description:** Left ventricular ejection fraction. • **Range:** 25-80.

	• **Name:** LBBB. • **Type:** int. • **Unit:** None. • **Description:** Presence of left bundle branch block on electrocardiogram. • **Range:** 0 (no), 1 (yes).
	• **Name:** Abnormal_Q_wave. • **Type:** int. • **Unit:** None. • **Description:** Presence of an abnormal Q wave on electrocardiogram. • **Range:** 0 (no), 1 (yes).
	• **Name:** 2_risk_factors_for_sudden_death. • **Type:** int. • **Unit:** None. • **Description:** Presence of >= 2 risk factors for sudden death. • **Range:** 0 (no), 1 (yes).
	• **Name:** t. • **Type:** float. • **Unit:** Year. • **Description:** Time to event for which to evaluate the model. • **Range:** 0-5 years.
Outputs of a function, model, or data collection protocols	• **Name:** survival_probability. • **Type:** float. • **Unit:** none. • **Description:** Event-free survival where the event is cardiovascular death. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of event-free relative survival in end-stage hypertrophic cardiomyopathy (HCM) patients, where the defined event is cardiovascular death. This is complementary to predicting cardiovascular mortality in patients with end-stage HCM. The function uses Cox Regression model without time-dependent covariates.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:4733774.
References to labels or annotations (if applicable)	See PMC:4733774.
References to computational methods (if applicable)	The model employs multivariate survival regression to estimate the probability of event-free survival.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later.

	• **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 99 end-stage HCM patients from a total of 1844 consecutive HCM patients from April 2002 to November 2013 at Fuwai Hospital, aged 52 ± 16 years old at entry. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Event-free relative survival where the event is all-cause death.
Methodological limitations	The model is based on data from a specific patient cohort and may not generalise well to other populations. The predictions are dependent on the accuracy of covariates. The model assumes proportional hazards over time, meaning the effect of covariates on the hazard is constant, which may not hold true in real-world settings and can lead to biased results if violated.
Example of use (optional)	See Appendix C.
Additional Information	After validation, the model can potentially be used in a clinical setting to predict cardiovascular mortality. This can help clinicians identify high-risk patients, enabling timely interventions and personalised management strategies. This can also support shared decision-making around therapies like ICD implantation or advanced monitoring. Additionally, the model may be used as a feature for predicting other HCM outcomes. In its current form, this unvalidated model can be used for research and development and is not intended for medical use.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/

	Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.5. Information given by: Orchard, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_SCA_ICD_w303mA.
Type of a function, model, or data collection protocol	Survival regression.
Short description of a function, model, or data collection protocol	Prediction of event-free relative survival where the event is defined as Sudden Cardiac Arrest in hypertrophic cardiomyopathy (HCM) patients who underwent implantable cardioverter defibrillator (ICD) implantation.
Inputs of a function or model	• **Name:** Sudden_Death_risk_factors. • **Type:** int. • **Unit:** None. • **Description:** Factors associated with the risk of sudden cardiac death in hypertrophic cardiomyopathy (HCM). • **Range:** 0 (no), 1 (yes).
	• **Name:** T_wave_Amplitude_Lead_V2_ • **Type:** float. • **Unit:** mV. • **Description:** Amplitude of T wave in lead V2 on electrocardiogram per 0.1 mV change. • **Range:** 0.1-1.0 mV.
	• **Name:** t. • **Type:** float. • **Unit:** Year. • **Description:** Time to event for which to evaluate the model. • **Range:** 0-5 years.
Outputs of a function, model, or data collection protocols	• **Name:** survival_probability. • **Type:** float. • **Unit:** none. • **Description:** Event-free survival where the event is sudden cardiac arrest. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of event-free relative survival where the event is defined as sudden cardiac arrest in HCM patients who underwent ICD implantation. The function uses Cox Regression model without time-dependent covariates.
Function, model, or protocol version (if applicable)	Ver 0.01.

References to data sources (if applicable)	See PMC:5384475.
References to labels or annotations (if applicable)	See PMC:5384475.
References to computational methods (if applicable)	The model employs multivariate survival regression to estimate the probability of event-free survival.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 52 patients with HCM who underwent implantable cardioverter defibrillator (ICD) implantation. • **Intervention:** ICD. • **Comparison:** None. • **Outcome:** Event-free relative survival where the event is sudden cardiac arrest.
Methodological limitations	The model is based on data from a specific patient cohort and may not generalise well to other populations. The predictions are dependent on the accuracy of covariates. The model assumes proportional hazards over time, meaning the effect of covariates on the hazard is constant, which may not hold true in real-world settings and can lead to biased results if violated.
Example of use (optional)	See Appendix C.
Additional Information	Once validated, the model could be used in clinical settings to predict the risk of sudden cardiac arrest (SCA) in patients who underwent ICD therapy. This would assist clinicians in identifying high-risk patients, facilitating timely interventions and personalised management strategies. It also supports shared decision-making regarding preventive treatments. Additionally, the model may be used as a feature for predicting other cardiovascular outcomes. In its current unvalidated state, it is intended solely for research and development and is not suitable for medical use.

Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.6. Information given by: Orchard, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_SCD_t0F9dg.
Type of a function, model, or data collection protocol	Survival regression.
Short description of a function, model, or data collection protocol	Prediction of sudden cardiac death (SCD) in patients with HCM in the Kochi RYOMA Study who are followed as part of the registry.
Inputs of a function or model	• **Name:** Male_sex. • **Type:** int. • **Unit:** None. • **Description:** Male sex • **Range:** 0 (no), 1 (yes).
	• **Name:** Age_at_diagnosis. • **Type:** float. • **Unit:** year. • **Description:** Age at diagnosis. • **Range:** 18-100.
	• **Name:** Family_history_of_HCM. • **Type:** int. • **Unit:** None. • **Description:** Family history of hypertrophic cardiomyopathy. • **Range:** 0 (no), 1 (yes).
	• **Name:** Family_history_of_SCD. • **Type:** int. • **Unit:** None. • **Description:** Family history of sudden cardiac death. • **Range:** 0 (no), 1 (yes).
	• **Name:** Presence_of_syncope. • **Type:** int. • **Unit:** None. • **Description:** Presence of syncope. • **Range:** 0 (no), 1 (yes).
	• **Name:** NSVT. • **Type:** int. • **Unit:** None.

	• **Description:** Presence of non-sustained ventricular tachycardia. • **Range:** 0 (no), 1 (yes).
	• **Name:** Maximum_LVWT. • **Type:** float. • **Unit:** mm. • **Description:** Maximum left ventricular wall thickness. • **Range:** 6-11.
	• **Name: Left_atrial_diameter.** • **Type: float.** • **Unit: mm.** • **Description: Left atrial diameter.** • **Range:** 29-40.
	• **Name:** LVOT_obstruction_at_registration. • **Type:** int. • **Unit:** None. • **Description: Presence of left ventricular outflow tract (LVOT) obstruction at registration.** • **Range:** 0 (no), 1 (yes).
	• **Name:** D_HCM_at_registration_or_follow_up. • **Type:** int. • **Unit:** None. • **Description:** Dilated phase of hypertrophic cardiomyopathy at registration or follow-up. • **Range:** 0 (no), 1 (yes)
	• **Name:** t. • **Type:** float. • **Unit:** Year. • **Description:** Time to event for which to evaluate the model. • **Range:** 0-5 years.
Outputs of a function, model, or data collection protocols	• **Name:** survival_probability. • **Type:** float. • **Unit:** none. • **Description:** Event-free survival where the event is cardiovascular death. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of sudden cardiac death (SCD) in 293 patients with HCM in the Kochi RYOMA Study who are followed as part of the registry. The function uses Cox Regression model without time-dependent covariates.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:7819656.
References to labels or annotations (if applicable)	See PMC:7819656.
References to computational methods (if applicable)	The model employs multivariate survival regression to estimate the probability of event-free survival.

Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 293 patients with HCM in the Kochi RYOMA Study who are followed as part of the registry. The mean (+-SD) age at registration and diagnosis was 63+-14 and 56+-16 years, respectively. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Event-free relative survival where the event is sudden cardiac death.
Methodological limitations	The model is based on data from a specific patient cohort and may not generalise well to other populations. The predictions are dependent on the accuracy of covariates. The model assumes proportional hazards over time, meaning the effect of covariates on the hazard is constant, which may not hold true in real-world settings and can lead to biased results if violated.
Example of use (optional)	See Appendix C.
Additional Information	After validation, the model can potentially be used in a clinical setting to predict cardiovascular mortality. This can help clinicians identify high-risk patients, enabling timely interventions and personalised management strategies. This can also support shared decision-making around therapies like ICD implantation or advanced monitoring. Additionally, the model may be used as a feature for predicting other HCM outcomes. In its current form, this unvalidated model can be used for research and development and is not intended for medical use.

Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.7. Information given by: Rossiter, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_RV_AllCause_vmD9Hg.
Type of a function, model, or data collection protocol	Logistic Regression.
Short description of a function, model, or data collection protocol	Prediction of right ventricular (RV) involvement using echocardiographic parameters.
Inputs of a function or model	• **Name:** LV_GLS. • **Type:** float. • **Unit:** None. • **Description:** Global longitudinal strain of the left ventricle. • **Range**: -35% to -5%.
	• **Name:** RV_free_wall_LS. • **Type:** float. • **Unit:** None. • **Description:** Longitudinal strain of the right ventricular free wall. • **Range:** -30% to -5%.
	• **Name:** RV_thickness. • **Type:** float. • **Unit:** mm. • **Description:** Thickness of the right ventricle. • **Range:** 1 to 10 mm.
	• **Name:** LV_maximal_thickness. • **Type:** float. • **Unit:** mm. • **Description:** Maximal thickness of the left ventricle. • **Range:** 6 to 30 mm.
Outputs of a function, model, or data collection protocols	• **Name:** risk probability. • **Type:** float. • **Unit:** none. • **Description:** Probability of the presence of RV involvement (impairment or changes in the right ventricle's structure or function). • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of right ventricular (RV) involvement using echocardiographic parameters, with improved accuracy by combining RV GLS, RV free wall LS, and RV wall thickness. RV involvement refers to structural or functional impairments

	in the right ventricle, detectable through imaging techniques like echocardiography, potentially indicating underlying heart conditions.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:7736330.
References to labels or annotations (if applicable)	See PMC:7736330.
References to computational methods (if applicable)	The model employs logistic regression to estimate the probability of RV involvement.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 256 patients with HCM who underwent both cardiac magnetic resonance (CMR) imaging and transthoracic echocardiography within 6 months of each other. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Probability of RV involvement.
Methodological limitations	The model may not generalise well to other populations and assumes linearity between predictors and log odds which might not be valid.
Example of use (optional)	See Appendix C.
Additional Information	After validation, the model can be used to identify high-risk patients for RV involvement, facilitate timely interventions and personalised treatments, support clinical decision-making, and enhance monitoring for conditions affecting the right ventricle. Additionally, the model may be used as a feature for predicting other HCM outcomes. In its current

	form, this unvalidated model can be used for research and development and is not intended for medical use.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.8. Information given by: Rossiter, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_LGE_hZnaDA.
Type of a function, model, or data collection protocol	Logistic Regression.
Short description of a function, model, or data collection protocol	Prediction of presence of Late Gadolinium Enhancement (LGE) as identified by contrast-enhanced cardiac magnetic resonance in patients with hypertrophic obstructive cardiomyopathy.
Inputs of a function or model	• **Name:** Calcium_antagonists. • **Type:** int. • **Unit:** None. • **Description:** Indicator for the use of calcium antagonists. • **Range:** 0 (no), 1 (yes).
	• **Name:** Resting_LVOTG. • **Type:** float. • **Unit:** mmHg. • **Description:** Resting left ventricular outflow tract gradient. • **Range:** 0-100 mmHg.
	• **Name:** MWT. • **Type:** float. • **Unit:** mm. • **Description:** Maximum wall thickness. • **Range:** 6-30 mm.
	• **Name:** LVEF. • **Type:** float. • **Unit:** None. • **Description:** Left ventricular ejection fraction. • **Range:** 5%-80%.
	• **Name:** Log_cTnI. • **Type:** float. • **Unit:** ng/mL.

	• **Description:** Logarithm of cardiac troponin I levels. • **Range:** -1 to 2 (corresponding to 0.1 to 100 ng/mL)
Outputs of a function, model, or data collection protocols	• **Name:** presence_of_LGE. • **Type:** float. • **Unit:** none. • **Description:** Relative risk of presence of Late Gadolinium Enhancement as identified by contrast-enhanced cardiac magnetic resonance. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of presence of Late Gadolinium Enhancement LGE in patients with hypertrophic obstructive cardiomyopathy using logistic regression. The model uses medication information, electrocardiographic variables, and a blood test that measures the levels of troponin proteins.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:4713160.
References to labels or annotations (if applicable)	See PMC:4713160.
References to computational methods (if applicable)	The model employs logistic regression to estimate the odds of the presence of Late Gadolinium Enhancement (LGE).
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** Patients with hypertrophic obstructive cardiomyopathy (HOCM) (n=163; mean age 47.2 ± 10.8 years; 38.7% females). • **Intervention:** None. • **Comparison:** None. • **Outcome:** Relative risk of presence of Late Gadolinium Enhancement.

Methodological limitations	The model may not generalise well to other populations and assumes linearity between predictors and log odds which might not be valid.
Example of use (optional)	See Appendix C.
Additional Information	Once validated, the model can be used clinically to identify high-risk patients likely to exhibit late gadolinium enhancement (LGE), enabling timely interventions and personalised treatment strategies. It supports clinical decision-making and enhances monitoring for cardiac conditions, helping clinicians better understand disease progression and manage patients accordingly. LGE is important because it can indicate myocardial fibrosis or scarring, which are linked to adverse cardiac events. Additionally, the model may serve as a feature for predicting other outcomes associated with cardiac diseases. In its current unvalidated form, it is intended solely for research and development and is not suitable for medical use.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.9. Information given by: Chen, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_Mortality_gA8rVQ.
Type of a function, model, or data collection protocol	Survival regression.
Short description of a function, model, or data collection protocol	Multivariate prediction of mortality (where mortality includes sudden cardiac death) in hypertrophic cardiomyopathy (HCM) patients.
Inputs of a function or model	• **Name:** LVEF. • **Type:** float. • **Unit:** None. • **Description:** Left ventricular ejection fraction. • **Range:** 10-90.
	• **Name:** Loop_recorder. • **Type:** int. • **Unit:** None. • **Description:** Indicator for the presence of an implantable loop recorder (ILR). • **Range:** 0 (no), 1 (yes).

	• **Name:** Chronic_heart_failure. • **Type:** int. • **Unit:** None. • **Description:** History of chronic heart failure. • **Range:** 0 (no), 1 (yes).
	• **Name:** NYHA_functional_class_at_last_visit. • **Type:** int. • **Unit:** None. • **Description:** New York Heart Association functional classification at last visit. • **Range:** 0 (I), 1(II), 2(III), 3(IV).
	• **Name:** t. • **Type:** float. • **Unit:** Year. • **Description:** Time to event for which to evaluate the model. • **Range:** 0-50 years.
Outputs of a function, model, or data collection protocols	• **Name:** survival_probability. • **Type:** float. • **Unit:** none. • **Description:** Event-free relative survival where the event is mortality. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Multivariate prediction of mortality (where mortality includes sudden cardiac death) in hypertrophic cardiomyopathy (HCM) patients. The function uses Cox Regression model without time-dependent covariates.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:5101433.
References to labels or annotations (if applicable)	See PMC:5101433.
References to computational methods (if applicable)	The model employs multivariate survival regression to estimate the probability of event-free survival.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).

System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 43 consecutive patients [mean age 38.5 +- 14.3 years; 25 (58.1%) male; and 13 (30.2%) black African] were prospectively. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Event-free relative survival where the event is mortality.
Methodological limitations	The model is based on data from a specific patient cohort and may not generalise well to other populations. The predictions are dependent on the accuracy of covariates. The model assumes proportional hazards over time, which may not hold true in real-world settings and can lead to biased results if violated.
Example of use (optional)	TBD.
Additional Information	After validation, the model can potentially be used in a clinical setting to predict mortality. This can help clinicians identify high-risk patients, enabling timely interventions and personalised management strategies. Additionally, the model may be used as a feature for predicting other HCM outcomes. In its current form, this unvalidated model can be used for research and development and is not intended for medical use.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.10. Information given by: Chen, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_VentricularArrhythmia_jpnBvQ.
Type of a function, model, or data collection protocol	Logistic Regression.
Short description of a function, model, or data collection protocol	Prediction of event-free survival, where the event is defined as the occurrence of ventricular arrhythmia (excluding sudden cardiac death). Ventricular events are defined as sustained and non-sustained ventricular tachycardias.

Inputs of a function or model	• **Name:** Fibrosis. • **Type:** int. • **Unit:** None. • **Description:** Presence of fibrosis in the myocardium. • **Range:** 0 (no), 1 (yes).
Outputs of a function, model, or data collection protocols	• **Name:** relative_risk. • **Type:** float. • **Unit:** none. • **Description:** Ventricular arrhythmia excluding sudden cardiac death. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Prediction of event-free survival, where the event is defined as the occurrence of ventricular arrhythmia, excluding sudden cardiac death. Ventricular events are defined as sustained and non-sustained ventricular tachycardias. The function uses a Logistic Regression classifier.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:7076262.
References to labels or annotations (if applicable)	See PMC:7076262.
References to computational methods (if applicable)	The model employs logistic regression to estimate the odds ratio for event-free survival prediction.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 91 patients with HCM referred for CMR on a 1.5T MR imaging system. • **Intervention:** None. • **Comparison:** None.

	• **Outcome:** Relative risk of occurrence of ventricular arrhythmia.
Methodological limitations	Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. If this assumption is violated, the model may produce inaccurate predictions.
Example of use (optional)	TBD.
Additional Information	After validation, the model could potentially be used in a clinical setting to predict ventricular arrhythmia risk. This can help clinicians identify high-risk patients, enabling timely interventions and personalised management strategies. The model may also be useful for predicting other cardiovascular outcomes. Currently, it is intended for research and development purposes.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.11. Information given by: Chen, Orchard, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_DiastolicDysfunction_Troponin_e7kJyw.
Type of a function, model, or data collection protocol	Linear Regression.
Short description of a function, model, or data collection protocol	Prediction of high sensitivity cardiac troponin T using imaging markers of diastolic dysfunction obtained via multivariable linear regression analysis.
Inputs of a function or model	• **Name:** SCD_Risk_Score____. • **Type:** float. • **Unit:** None. • **Description:** Percentage sudden cardiac death risk score. • **Range:** 0-100.
	• **Name:** LA_diameter__mm_. • **Type:** float. • **Unit:** mm. • **Description:** Diameter of the left atrium. • **Range:** 20-70.
	• **Name:** Resting_LVOT_flow_gradient__ln___. • **Type:** float.

	• **Unit:** None. • **Description:** Logarithmised resting left ventricular outflow tract flow gradient. • **Range:** -3-6.
	• **Name:** Provoked_LVOT_flow_gradient__ln___. • **Type:** float. • **Unit:** None. • **Description:** Logarithmised provoked left ventricular outflow tract flow gradient. • **Range:** -3-6.
	• **Name:** E_E__mean. • **Type:** float. • **Unit:** None. • **Description:** Mean ratio of peak early transmitral filling velocity to early mitral annulus velocity. • **Range:** 0-30.
	• **Name:** E_E__septal. • **Type:** float. • **Unit:** None. • **Description:** Ratio of peak early transmitral filling velocity to early mitral annulus velocity at the septal annulus. • **Range:** 0-30.
	• **Name:** E_E__lateral. • **Type:** float. • **Unit:** None. • **Description:** Ratio of peak early transmitral filling velocity to early mitral annulus velocity at the lateral annulus. • **Range:** 0-30.
	• **Name:** IVRT_septal. • **Type:** float. • **Unit:** ms. • **Description:** Isovolumetric relaxation time at the septal annulus. • **Range:** 50-150.
	• **Name:** IVRT_lateral. • **Type:** float. • **Unit:** ms. • **Description:** Isovolumetric relaxation time at the lateral annulus. • **Range:** 50-150.
	• **Name:** SW_thickness__mm_. • **Type:** float. • **Unit:** mm.

	• **Description:** Thickness of the septal wall. • **Range:** 5-50.
	• **Name:** LGE_size__g_. • **Type:** float. • **Unit:** g. • **Description:** Size of late gadolinium enhancement. • **Range:** 0-100.
Outputs of a function, model, or data collection protocols	• **Name:** troponin_T__measurement. • **Type:** float. • **Unit:** ng/ml. • **Description:** high sensitivity cardiac troponin T. • **Range:** 0-0.99.
An extended description of a function, model, or data collection protocol	Prediction of high sensitivity cardiac troponin T (hs-cTnT) in hypertrophic cardiomyopathy patients using late gadolinium enhancement (LGE) and other imaging biomarkers of diastolic dysfunction using multivariate linear regression.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:9410236.
References to labels or annotations (if applicable)	See PMC:9410236.
References to computational methods (if applicable)	The model employs multivariate linear regression to estimate high-sensitivity cardiac troponin T.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 366 HCM patients who underwent transthoracic echocardiography (TTE), 218 of whom also obtained cardiovascular magnetic resonance (CMR) to assess focal myocardial fibrosis by LGE.

	• **Intervention:** None. • **Comparison:** None. • **Outcome:** high-sensitivity cardiac troponin T.
Methodological limitations	The model is based on a specific cohort and may not generalise well to other populations. Linear regression assumes a linear relationship between inputs and the output, which may not capture complex interactions.
Example of use (optional)	TBD.
Additional Information	After validation, the model could potentially be used in clinical settings to help identify and manage diastolic dysfunction in HCM patients, and/or to refer patients for high-sensitivity cardiac troponin T blood testing. At present, the model is intended for research and development purposes only and is not approved for clinical use.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/ Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Table WP6.12. Information given by: Chen, Orchard, Agakov [Pharmatics] [WP6]	
Field	**Explanation**
Name of a function, model, or data collection protocol	HCM_IschemicStroke_W8YN6A.
Type of a function, model, or data collection protocol	Survival regression.
Short description of a function, model, or data collection protocol	Cox regression for prediction of event-free relative survival in hypertrophic cardiomyopathy patients without documented atrial fibrillation, where the event is defined as ischemic stroke. The model is used to provide a relative hazard of ischemic stroke.
Inputs of a function or model	• **Name:** Age_____65_years. • **Type:** int. • **Unit:** None. • **Description:** Indicator for age >= 65 years. • **Range:** 0 (no), 1 (yes).
	• **Name:** Sex__Female. • **Type:** int. • **Unit:** None. • **Description:** Female sex. • **Range:** 0 (no), 1 (yes).
	• **Name:** Hypertension.

	• **Type:** int. • **Unit:** None. • **Description:** Diagnosis of hypertension. • Range: 0 (no), 1 (yes).
	• **Name:** Diabetes_mellitus. • **Type:** int. • **Unit:** None. • **Description:** Diagnosis of diabetes mellitus. • **Range:** 0 (no), 1 (yes).
	• **Name:** Chronic_heart_failure. • **Type:** int. • **Unit:** None. • **Description:** Diagnosis of chronic heart failure. • **Range:** 0 (no), 1 (yes).
	• **Name:** Vascular_disease__. • **Type:** int. • **Unit:** None. • **Description:** Presence of vascular disease (including myocardial infarction and peripheral artery disease). • **Range:** 0 (no), 1 (yes).
	• **Name:** t. • **Type:** float. • **Unit:** year. • **Description:** Time to event for which to evaluate the model. • **Range:** 1-5 years.
Outputs of a function, model, or data collection protocols	• **Name:** survival_probability. • **Type:** float. • **Unit:** none. • **Description:** Event-free relative survival where the event is ischemic stroke. • **Range:** [0, 1].
An extended description of a function, model, or data collection protocol	Cox regression for prediction of event-free relative survival in hypertrophic cardiomyopathy patients without documented atrial fibrillation, where the event is defined as ischemic stroke. The model uses a Cox Regression model without time-dependent covariates.
Function, model, or protocol version (if applicable)	Ver 0.01.
References to data sources (if applicable)	See PMC:9499955.
References to labels or annotations (if applicable)	See PMC:9499955.

References to computational methods (if applicable)	The model employs multivariate survival regression to estimate the probability of event-free survival.
Software (if applicable)	Details about the software environment required to run the function/model: • **Programming language:** python 3.8 or later. • **Required packages/Libraries:** numpy 1.19.0 or later, scikit-learn 0.24.0 or later. • **Configurations:** none. • **Environment variables:** none. • **Licensing agreements:** custom proprietary license. • **Installation instructions:** no installation required (API keys will be provided for the cloud access).
System and computing requirements (if applicable)	• **Operating system:** Windows or Linux. • **Hardware:** minimal requirements TBD. • **Batch mode:** batch or online use. • **Execution:** single CPU core. • **Running time:** TBD.
Intended use	Clinical research or clinical decision support.
Detailed clinical use (optional)	• **Population:** 8,328 HCM patients without documented AF and 1:2 propensity score-matched 16,656 non-HCM controls between 2010 and 2016. • **Intervention:** None. • **Comparison:** None. • **Outcome:** Event-free relative survival where the event is ischemic stroke.
Methodological limitations	The model is based on a specific patient cohort and may not generalise well to other populations. The predictions are dependent on the accuracy of covariates. The model assumes proportional hazards over time, meaning the effect of covariates on the hazard is constant, which may not hold true in real-world settings and can lead to biased results if violated.
Example of use (optional)	TBD.
Additional Information	After validation, the model can potentially be used in a clinical setting to predict ischemic stroke. This can help clinicians identify high-risk patients, enabling timely interventions and personalised management strategies. This can also support shared decision-making around therapies like anticoagulation. Additionally, the model may be used as a feature for predicting other HCM outcomes. In its current form, this unvalidated model can be used for research and development and is not intended for medical use.
Reference to evaluation	Not yet available.
Support team contact details	SMASH-HCM support for general enquiries: https://smash-hcm.eu/contact/

	Pharmatics MedAI technical support for technical enquiries: support@pharmaticsltd.com

Appendix C: Examples of MedAI-HCM API Calls

Figure 12 illustrates several command-line calls to a local version of the initial SMASH-HCM API for a subset of survival regression functions, outputting probabilities of cardiovascular or all-cause survival for individuals with diagnosed HCM, or time to sudden cardiac arrest in HCM patients who underwent ICD interventions. Figure 13 shows exemplar calls to binary classification functions that predict the risk of all-cause mortality and the presence of Late Gadolinium Enhancement (LGE), as identified by contrast-enhanced cardiac magnetic resonance. The JSON files contain patient-specific information, as illustrated in Figure 14.

```
(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
uv_HCM_VA_WQjTRA.json
{"result":[0.9593221770706931],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_VA_g0yDCg.json
{"result":[0.8795741307748108],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_AllCauseDeath_giznrw.json
{"result":[0.8708129522800446],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_CVDeath_bSKXXA.json
{"result":[0.1943309327609585],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
uv_HCM_CVDeath_3F2rng.json
{"result":[0.0262522106272265],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_ICD_w303mA.json
{"result":[0.5626877926146632],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_SCD_t0F9dg.json
{"result":[0.639765427291095],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_SCD_u4ZZXQ.json
{"result":[0.9565069535596003],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_HF_u0slqw.json
{"result":[0.44257819652557373],"jsonrpc":"2.0","id":0}

(chcm) P:\off\pharmatics\MedAI_instances\hcm\examples>python run_example.py --json
HCM_SCD_bSN2GQ.json
{"result":[0.5903743396552859],"jsonrpc":"2.0","id":0}
```

Figure 12 Examples of command-line calls to a subset of survival regression functions included in the initial SMASH-HCM API. To demonstrate local execution, a Python script uses patient-specific data from a JSON file to perform the analysis.

```
(venv)cody@fedora:~/Documents/code/MedAI_common$ python run_example.py --json
/home/cody/Documents/code/MedAI_hcm/examples/HCM_AllCause_vmD9Hg.json
{"result":[0.5447891928908387],"jsonrpc":"2.0","id":0}
{"result":[0.5638477262762027],"jsonrpc":"2.0","id":1}
{"result":[0.6193477504516473],"jsonrpc":"2.0","id":2}
{"result":[0.5520726890943916],"jsonrpc":"2.0","id":3}
(venv) cody@fedora:~/Documents/code/MedAI_common$ python run_example.py --json
/home/cody/Documents/code/MedAI_hcm/examples/HCM_LGE_hznaDA.json
{"result":[6.57147655146505e-5],"jsonrpc":"2.0","id":0}
```

Figure 13 Examples of command-line calls to a subset of binary classification functions included in the initial SMASH-HCM API.

```
[
  {
    "method": "HCM_AllCauseDeath_giznrw",
    "params": [
      {
        "$MaleSex": 1,
        "$Age": {"value": 25.47, "unit": "year"},
        "$NYHAClass3Or4": 1,
        "$AtrialFibrillation": 1,
        "$LVEF": 67.56,
        "$LBBB": 1,
        "$AbnormalQWave": 1,
        "$>=2RiskFactorsForSuddenDeath": 0,
        "$TimeToEvent": {"value": 4.5, "unit": "year"}
      }
    ]
  }
]
```


Figure 14 Example of patient-specific data for evaluating all-cause mortality within 4.5 years from the point of inference for a 25-year-old male with diagnosed HCM, NYHA class 4, atrial fibrillation, abnormal Q-wave, LBBB, but without two or more risk factors for sudden death.