PRECISE4Q

PREDICTIVE MODELLING IN STROKE

DELIVERABLE

Project Acronym: **Precise4Q**

Grant Agreement number: **777107**

Project Title: **Personalised Medicine by Predictive Modelling in Stroke for better Quality of Life**

D4.4 – White paper on hybrid model fusion strategies

Revision: 0.12

Authors and Contributors	Main author: Gunnar Cedersund (LIU); Contributors: Vince Madai (CUB); Tilda Herrgårdh (LIU), Attia Fatima (DIT), John Kelleher (DIT)			
Responsible Author	Gunnar Cedersund		**Email**	gunnar.cedersund@liu.se
	Beneficiary	LIU	**Phone**	+46-702-512323

Project co-funded by the European Commission within **H2020-SC1-2016-2017/SC1-PM-17-2017**		
Dissemination Level		
PU	Public, fully open	**X**
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

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

Revision History

Revision	Date	Author	Organisation	Description
0.1	5.2.2019	Gunnar Cedersund	LIU	First bullet points version, and vs 1 of 5 key figures. Followed by first meeting with all contributors
0.2	5.2.2019	John Kelleher and Vince Madai	CUB and DIT	Comments on vs 0.1
0.3	12.2.2019	Gunnar Cedersund	LIU	Version with introduction written out as full text
0.4	12.2.2019	all contributors	LIU, DIT, CUB	Comments on the version 0.3
0.5	16.2.2019	Attia Fatima	CUB	Text and references regarding sub-chapter on machine learning and bioinformatics
0.6	17.2.2019	Gunnar Cedersund	LIU	Half of the text is now written
0.7	25.2.2019	Gunnar Cedersund	LIU	A complete version is now in place, with all text and figures present
0.8	27.2.2019	Tilda Herrgårdh	LIU	Polishing of text, integration of comments
0.9	27.2.2019	Vince Madai	CUB	Revision of document
0.10	28.2.2019	Gunnar Cedersund	LIU	Final revisions and polishing of all the text
0.11	28.2.2019	Tilda Herrgårdh	LIU	Final layout issues in offline version. Send to upload.
0.12	28.2.2019	Dage Särg	UTARTU	Reviewed

Date of delivery	Contractual:	28.2.2019	Actual:	28.2.2019
Status	final X /draft ☐			

Abstract (for dissemination)	This deliverable describes the approaches to be used in PRECISE4Q regarding the hybrid modelling. In other words, herein, we specify how the mechanistic, bioinformatics, and machine learning models will be combined. First the three fields are described, including state-of-the-art methodology and some

	key models. Second, we go through how the combination will be done regarding calculation of risk factors. This is done in a two step fashion. In the first step, a blended hybrid modelling approach is done, using methodologies from nonlinear mixed-effects modelling, which allows for distributions of parameters and the addition of covariates into mechanistic models. In the second step, the output of those models, and the output of bioinformatics network models, are inputted into machine learning models, in a sequential hybrid modelling approach. The third part of the deliverable describes our approach to hybrid modelling during the simulation of scenarios. This is an 8 step approach, that switches several times between mechanistic modelling and machine learning. The key idea is to use the mechanistic modelling for the simulations forward in time, and to use the machine learning modelling to extend from a few measured or simulated values, to all other values. The described approaches will be applicable to all phases of stroke that will be modelled using hybrid modelling, which primarily means the prevention and acute phases.
Keywords	mechanistic modelling, multilevel modelling, hybrid modelling, bioinformatics, machine learning, stroke, prevention, acute treatment

Statement of originality

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

Table of Content

List of Figures

Executive Summary

PRECISE4Q deals with the development of new systems medicine approaches to help aid stroke patients, in all phases of disease: during prevention, acute treatment, recovery and reintegration. Work Package 4 (WP4) deals with the development of the computational models, which are based on the information and data from the previous WPs, which then is used in clinical studies in WP5.

One of the key ideas behind the modelling in WP4 is to use hybrid modelling. Hybrid models are models that combine two or more modelling approaches into a joint model. In this project, we will combine three types of models: machine-learning, bioinformatics, and mechanistic models. Each of these model types has their own respective strengths. Machine-learning is good because of its flexibility, because it can learn from data without the need for a mechanistic understanding, and because it can provide statistical assessments, such as risk scores. Machine-learning models often require large training data, where the same input and output have been measured in a large number of patients. Bioinformatics models are also developed to deal with large-scale data, but they are often large in another sense: the number of variables is large. Such data are often called omics data, which is a joint name for e.g. metabolomics (the measurement of all metabolites), proteomics (data for all proteins), genomics (all genes), etc. Omics data can be available for few, even one, patient, and bioinformatics methodologies use techniques such as network models to infer joint properties of the variables, such as clusters of co-regulated genes. These joint properties may serve as new biomarkers, which are not present in the original data. Such new biomarkers can also be obtained by the last modelling type considered herein: mechanistic models. With mechanistic models, one also uses networks, but then these networks are dynamic, usually described by ordinary differential equations, which are built to represent the available mechanistic knowledge for the system. The addition of this mechanistic knowledge adds new useful information to the data analysis problem, which means that mechanistic models can work also with small-scale data.

In a hybrid model that combines all of these three modelling approaches - machine learning, bioinformatics and mechanistic models - the strengths of the three approaches are combined. The mechanistic and physiological knowledge is made use of - and added onto - by the mechanistic modelling, omics data can be utilized by the bioinformatics modelling to construct new biomarkers, and statistical assessments such as risk scores and diagnoses can be calculated using the machine learning. Furthermore, using hybrid models one can also do simulations of scenarios, which can investigate e.g. how a patient would be expected to respond to a new treatment, such as a drug, a diet, or to exercise. Using hybrid models one can also both get risk scores and scenarios, and always ask the question *why* these risk scores and scenarios are obtained. This question is answered, by looking inside the model, to see which mechanisms and specific patient properties that caused the specific prediction.

However, even though these types of hybrid models have obvious advantages, quite few hybrid models have yet been developed in the systems medicine community. Some reviews and roadmaps are available, but few concrete examples exists, and none for the specific topic of stroke treatment, which is the goal of PRECISE4Q. For this reason, this deliverable, D4.4, outlines our strategy for how the hybrid modelling will be performed. There are two types of strategies outlined, one for calculations of risk scores, and one for scenario simulations.

For calculations of risk scores, we will use a two step approach (Fig 8). In the first step, a blended hybrid modelling approach, employing a methodology called nonlinear mixed-effects modelling will be used. This, together with bioinformatics analysis of the omics data, will be used to produce new biomarkers, which only are obtainable using a combination of the information in the data, and the information in the process understanding, implemented via the models. In the second step, these

new model-derived biomarkers are combined with the rest of the biomarkers, coming from the data directly, to produce an ultimate risk score calculation, or a diagnosis. This combined two-step approach is referred to as sequential hybrid modelling, and the second step, combining all available biomarkers is made using an appropriate machine learning methodology, such as a Bayesian Graphical Network, or deep learning.

For simulation of scenarios, we have outlined a new 8-step approach (Fig 9). In the first three steps, the data available for the patient is expanded to more data, using machine learning approaches or imputation, and available prior knowledge. This allows for the mechanistic model to be instantiated, and simulated, which are steps four and five. Thereafter, for each timepoint in the simulations, a similar machine learning or imputation approach as was done in step two can be used again, to infer reasonable updated values, at each time point, also for the omics data and for the other biomarkers. These inferred and updated values then go into the risk score hybrid model (Fig 8), which is used to also calculate a risk score for each time point.

In summary, this delivery outlines our strategies for how the hybrid modelling in PRECISE4Q will be done. This will allow us to move forward with the analysis of the data from WP2 and WP3, to create the computational models. Because of our novel hybrid modelling approaches, we will be able to both calculate statistical assessments such as diagnosis and risk scores, to make use of all data and all prior knowledge, and to simulate new scenarios. These possibilities will be useful in the different use case scenarios outlined in WP1 and D4.1-2, and will be implemented and tested in new clinical studies, performed in WP5. The presented approaches are general, and will be applicable to all cases where hybrid modelling is needed, which primarily will be for prevention (such as in Use Case 3) and acute treatment (Use Case 2-4).

1 Scope and Purpose

The goal of PRECISE4Q is to minimise the burden of stroke for the individual and for society. This will be done by creating data-driven predictive machine learning models, which enables – for the first time – personalised stroke treatment, addressing patient needs in all four stages: prevention, acute treatment, rehabilitation, and reintegration.

In some of the previous deliverables in other work packages (WP1, D.1.1- D1.3), we have already outlined some specific use cases and scenarios that will be the main focus of PRECISE4Q. Here in WP4, we take these general wishes, from the clinical perspective, convert them to more concrete and realistic goals from a modelling perspective, and deliver the actual mathematical models, which are then used in the later WPs. More specifically, in D4.1, we have refined our initial assessment and use cases from D1.1, to an updated list of use cases, and of risk, health and resilience factors. These factors, the models will take as inputs. Thereafter, in D4.2, we defined the targets that the models should be able to deliver as outputs. These will be the outputs of the models.

To be able to do such combinations of both i) statistical mappings from inputs (risk, health and resilience factors) to outputs (outcomes), and ii) to simulate scenarios of how these variables change over time, one needs to develop hybrid models. Such hybrid models combine mechanistic multilevel models, describing the underlying physiological and biochemical processes, with statistical machine-learning models. Such models are relatively novel in the scientific community. There are several approaches proposed in the literature, but few implemented for specific medical applications, and seemingly none of them for stroke. Therefore, to prepare for the actual work to follow, in this deliverable (D4.4), we will outline how this hybrid modelling will be done. In other words, ***in this deliverable we will describe how we will combine the mechanistic and machine learning models, to be able to accomplish the combination of both outcome prediction and the simulation of specific scenarios.***

To help the reader to see these dependencies between the relevant deliverables submitted at this point, here is a summary.

D1.1. Risk factors and outcome

This deliverable summarized the scientific status regarding risk factors and targets for each phase of stroke.

D1.2. Clinical challenges and needs

Here, we outlined the clinical needs for each stroke emphasizing the most common questions.

D1.3 Use Cases with inputs/outputs

This deliverable summarized the baseline use cases which have been explored in the literature so far.

D4.1 White paper on stroke risk, health and resilience factors

This deliverable will provide a) updated use cases as a basis for deliverables 4.1, 4.2 and 4.4. and b) provide an overview which risk, health and resilience factors we will explore in the modelling phase. A major point here is the point-of-view regarding potential interventions, i.e. the implementation of better clinical care through machine learning based clinical decision support

D4.2 Defining prediction targets for the models

Whereas 4.1 focuses on the features/factors which are relevant for the use cases in 4.2. we will summarize which targets we will predict in our modelling phase. Next to existing markers - as baseline - we will explore the development of new, complex Quality of life markers within P4Q.

D4.4. Patient Outcome Heterogeneous Model Fusion Strategies

A highly promising approach for the successful development of personalized clinical decision support tools is the use of combined mechanistic simulation and phenomenological machine learning tools. Here, - based on the use cases - we will explore the possible application of hybrid models for our modelling approaches.

The rest of the deliverable is structure as a review paper, and it will - after appropriate modifications and further developments - also be submitted for journal publication.

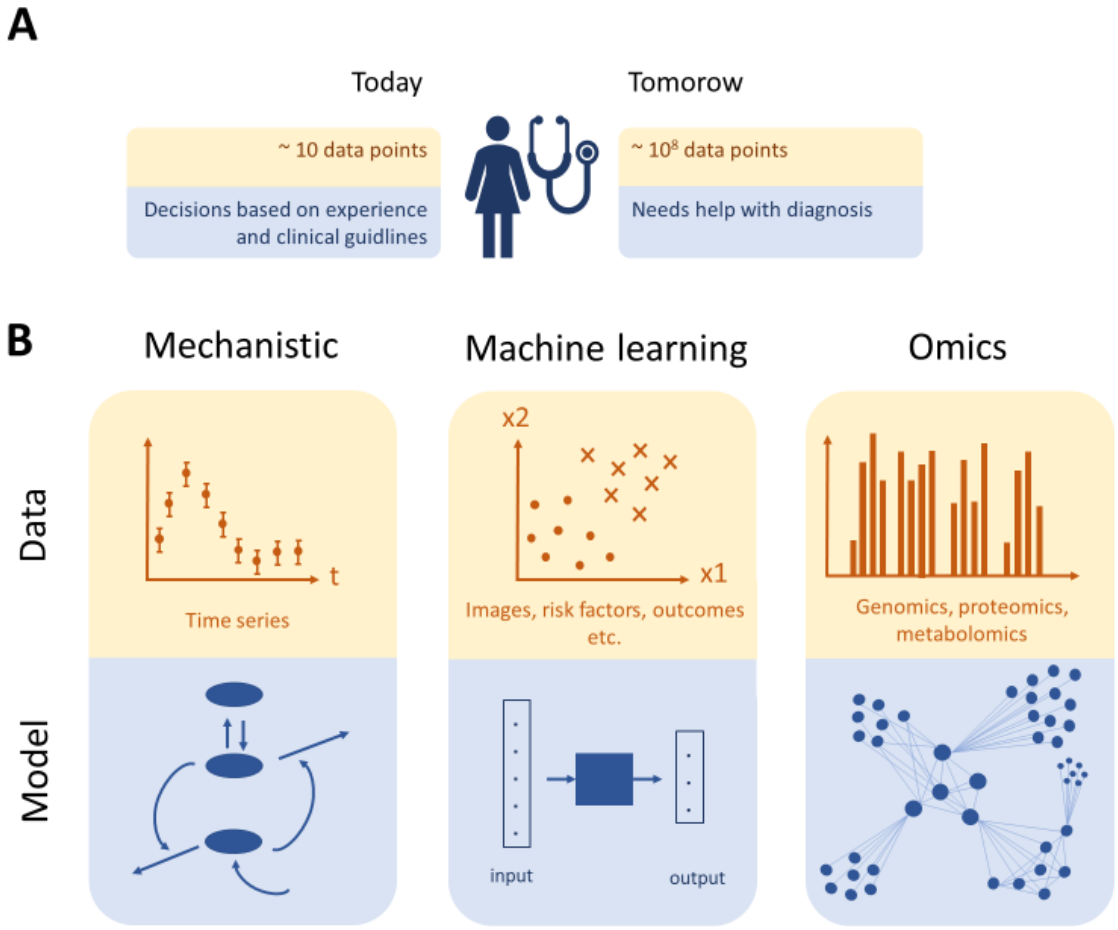

Figure 1: *The challenge of big data: (A) Today the physicians can manage by conventional inspection and reasoning around data, based on experience and clinical guidelines; tomorrow this will be no longer be a feasible solution. (B) Overview of the 3 main types.*

2 Background

The challenge of Big Data warrants the development of new Precision Medicine technologies

In the healthcare sector we have, as in many other parts of society, entered into the age of Big Data. Up until now, it has been sufficient for a physician who is about to diagnose a patient to look at a handful of biomarkers, do an interview with the patient, and then go to his/her experience and to clinical guidelines, to know what to do next (Fig 1A). This strategy is now no longer sufficient. With the advent of modern measurement and storage technologies, the amount of real world data available from a patient visit can easily be so large that they cannot even be inspected manually. Apart from that, we now have access to large clinical studies, with millions of similar patients examined and followed over time, which all serve as important background information. Finally, to add insult to injury, the patient may him/herself also have access to own smart sensor technology, which themselves generate huge amounts of data, of often unknown quality. To simply ignore all of that new data and knowledge, now available to a physician, is not an acceptable solution, and it will not be permitted long by society. Therefore, new technologies, which can make use of and integrate all of this data and knowledge are urgently needed.

The use of this vast amount of data for highly individualized predictions is called Precision Medicine. It is a form of health care that emerged in the past years that relies on data, algorithms, and precision molecular tools to offer individualized care for patients (Dzau and Ginsburg, 2016). It gives insight into mechanisms of disease, treatment and prevention. By treating the patient as an individual, the attending physician is able to consider variations in pathophysiology, genome, and anatomical variances. This has the potential to improve outcomes and to reduce healthcare costs. Precision medicine has for example already been successfully used in oncology, to find genetic mutations, and it is now considered for a variety of different fields (Hinman et al., 2017; Rostanski and Marshall, 2016). Its use in stroke is now emerging as more and more pathophysiological data becomes available. Stroke has a complex pathophysiology, comprising medical and environmental factors and is therefore a suitable candidate for precision medicine (Rostanski and Marshall, 2016). Different types of data like clinical and imaging data are available for ischemic stroke. Additionally, given its high prevalence, a lot of data is routinely acquired and can be made available. A precision medicine approach can therefore integrate this data and offer better treatment decision making and outcome prediction (Dzau and Ginsburg, 2016; Hinman et al., 2017).

Three types of data, and three corresponding types of modelling approaches

The available data come in different forms, and can be subdivided, e.g., according to the method by which they can be analysed (Fig 1B). One such example is such data that can be analysed using mechanistic modelling. Such data can e.g. be blood concentrations of glucose and insulin, which can be described by metabolic models based on ordinary differential equations (ODEs) describing their interplay on the whole-body level (Man et al., 2007; Nyman et al., 2011); it can be blood pressures and flows, which can be described by either zero-dimensional ODEs, or partial differential equations (PDEs) (Casas et al., 2018, 2017); and it can be intracellular concentrations of metabolites and proteins, which may be described by ODE models that are based on metabolic and signalling pathways (Brännmark et al., 2017, 2013; Nyman et al., 2014). Such data and modelling would

normally appear in fields such as systems biology, biomechanics, systems pharmacology, etc. Note that all of the examples above are of models that are relevant for stroke, and that all of these models in PRECISE4Q will be combined into a large inter-connected multi-level and multi-timescale model. Another type, which needs another set of analysis tools are those belonging to the fields of omics and bioinformatics. Omics is a word that jointly describes data that describes all of the available units in one go; some examples include, e.g. proteomics, which describe all proteins; metabolomics, which describes all metabolites; and genomics, which describes all genes or the expression of all genes. These types of data are usually analysed using more simple tools than ODEs, and are instead based on network analyzes, that try to identify clusters of genes that together may express a significant change, even when each gene individually is not significantly changed by themselves (Gustafsson et al., 2014; Kim and Tagkopoulos, 2018; Rappoport and Shamir, 2018). A final set of data and analysis approaches are known as machine learning or phenomenological models. Machine learning approaches are based on data with a given set of inputs (e.g. biomarkers, risk and resilience factors), and a given set of outputs (e.g. outcomes, such as if the patient will get a stroke within the next 4 years) (Kelleher et al., 2015).

In summary, it is therefore clear that there are different types of data, and different types of analysis tools. Let us consider these modelling technologies again, and this time identify the strengths and weaknesses of each approach.

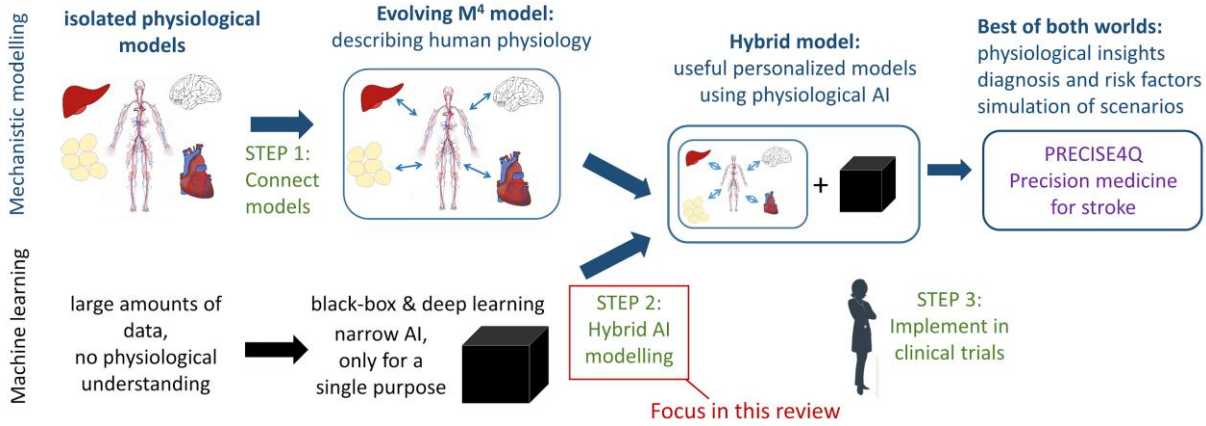

Figure 2: *Overview of the benefits of combining M4 models with machine learning: you can both get physiological insights, diagnosis and risk factors based on all data, do simulation of scenarios, and test treatments in silico. The overall goal of PRECISE4Q is to develop the models, combine them into hybrid models, and to do relevant clinical trials that allow for testing of the models. The focus of this review (D4.4) is on how to do the hybrid modelling.*

Phenomenological models, machine learning and bioinformatics

A key benefit of machine learning and bioinformatics approaches is that they do not require understanding of the processes involved, i.e. they do not need any type of prior knowledge. In other words, these methods are, by themselves, purely data-driven, and can generate e.g. a network or statistical mapping from inputs to outputs from the data directly. Such models are often referred to as black-box models (Fig 2, lower left). The drawbacks of these approaches are that they often require large number of datapoints (N>10 000 ideally) to be able to extract this information. This is

especially true for many machine learning approaches. Nevertheless, the ever increasing amount of features also leads to what is known as the “curse of dimensionality” which strongly limits our ability to extract relevant personalized information from Big Data, even using machine learning methodologies (Barbour, 2019). Another drawback with machine learning models is that they are hard to generalize, and that they do not make use of or add to the available physiological knowledge about the system. In other words, with the basic versions of these methods, it is hard to ask the question “why” a certain prediction is obtained. There are ongoing developments to help resolve this (Handelman et al., 2018; Medicine, 2018; VoosenJul. 6 et al., 2017). The perhaps most important such approach is the hybrid modelling approach, proposed herein. Finally, bioinformatics suffers from many of the same problems, but there are some key differences: in bioinformatics, the result can often be a module or a network, which has a biological interpretation, and some background knowledge typically is added as a *prior* to the inferral of these network models. Similarly, in bioinformatics the largest aspect of the data is that many variables are measured; the number of repeats is not necessarily as big as in a typical machine learning setting (Fig 1B).

Mechanistic multi-level, multi-timescale and multi-species modelling (M4-models)

The benefits and drawbacks of mechanistic models are almost exactly the inverse of that of machine learning models. Some of the most important benefits are that it is possible to input physiological knowledge; that one can work with small datasets, as long as clear changes can be seen in the data; and that one can always ask the question “why” a particular prediction is made. Another important benefit is that mechanistic models can be reused, and easily expanded in new contexts, as new data and applications become available. Some of the most important drawbacks of systems biology mechanistic modelling are that they themselves are usually not statistical in nature, i.e. they usually do not produce such things as risk scores, which is an inherently statistical and phenomenological property. Furthermore, most systems biology mechanistic models are usually mean value models, and advanced patient-specific modelling requires the expansion of purely mechanistic modelling to also include phenomenological covariates, as is done in nonlinear mixed-effects modelling (Jonsson et al., 2000; Karlsson et al., 2015). That is one of the hybrid modelling approaches discussed further below.

The need for hybrid modelling

This comparison between phenomenological machine-learning or bioinformatics models, on the one hand, and mechanistic multilevel modelling, on the other hand, quickly reveals that they have complementary strengths and weaknesses. The strongest possible models would therefore, in principle, be hybrid models that combines the two modelling approaches (Fig 2, Step 2). Some such hybrid models have been proposed for biomedical applications. One of those proposals came from the Discipulus network, which produced a roadmap for how “Digital Patients” - in silico representations of individuals - can be developed (“Digital Patient Roadmap,” n.d.). This roadmap outlines many useful ideas ranging from data integration and handling, to modelling approaches, and even to clinical applications. However, no concrete models were developed in that network, even though some existing non-hybrid models for stroke are mentioned in the roadmap. There are also some specific hybrid models developed for biological applications, summarized in e.g. these reviews (Doyle et al., 2013; Stéphanou and Volpert, 2016). From these reviews a couple of things become

clear i) hybrid models are still rare, but their incidence has been rapidly evolving over the last couple of years (Stéphanou and Volpert, 2016); ii) there is no consensus regarding nomenclature, but some basic options (Fig 3) for how the models can be combined are emerging; iii) there are no fully developed hybrid models that combines mechanistic multilevel models with machine learning and bioinformatics in a clinically useful way, and especially none for stroke.

Purpose and outline of review

For all of these reasons, it is therefore clear that there is a strong need to help mature the field of hybrid modelling. We herein contribute to this, by presenting a more concrete pathway forward with clear illustrations and comments regarding how hybrid models for stroke could be developed. The methodologies and strategies are completely general, but we primarily illustrate their usage for prevention of stroke, and for acute treatment. First follows a slightly more in-depth review of each of the ingoing modelling approaches, and a deeper look at some of the state-of-the-art models (Fig 4-7). Thereafter, a two-step approach for the calculation of risk scores (Fig 8), and an eight-step approach for simulation of scenarios (Fig 9) is introduced.

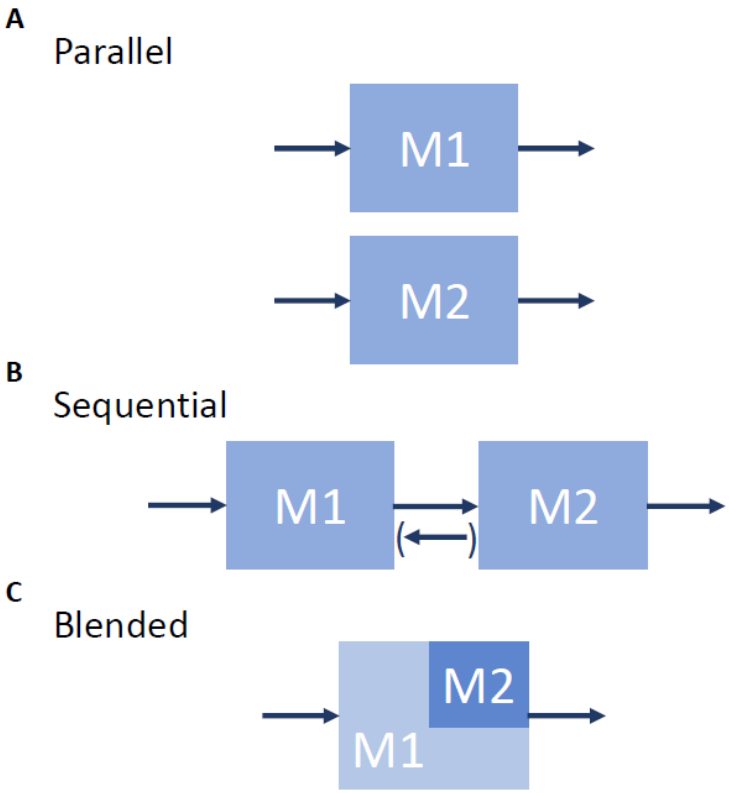

Figure 3: *Overview of the three main approaches to hybrid modelling most often adopted today. Note that the denomination of these approaches has not yet converged, and that other coexisting names for these and related approaches have been proposed (Stéphanou and Volpert, 2016). For instance, sequential modelling is sometimes referred to as iterative modelling or staged hybrid models, depending on if the feedback from M2 to M1 is included.*

3 Review of existing methodologies and concepts

3.1 Phenomenological models, machine learning, and bioinformatics

Phenomenological models and machine learning

Machine learning is a scientific discipline based on the idea that systems can learn from data (Lip et al., 2010). Machine learning approaches recognize patterns in data for the analytical prediction model building (Kelleher et al., 2015). Two main machine learning approaches are classified into two main classes: supervised and unsupervised learning. In the case of supervised learning, the desired outputs are given beforehand. Inputs are mapped to pre-defined outputs, and model training is done on those known outputs. Unsupervised learning involves searching for structure in the data (Alpaydin, 2009). Two state-of-the art machine learning approaches for predictive model generation in biomedicine are artificial neural networks (AAN) and support vector machines (SVM) (Warwick, 2004).

An ANN is constructed like, and functions similar to, a model of neuronal activity in living organisms (Wasserman, 1993). In other words, the model systems work as interconnected networks of units or nodes (“neurons”), the systems take input information to learn from it and based on complex relationship or connections give an output, and the networks adapt by making changes in their connections based on external or internal information. This property of neural networks helps to build dynamic system models describing the relation between inputs and outputs (Warwick, 2013).

The support vector machine (SVM) is a popular machine learning classifier in medicine, which was performance-wise the best general purpose algorithm prior to the advent of ANNs and also works quite well with limited amount of data. For model training, a set of data with known classification is given to the algorithms. This leads to building a model capable of assigning new data into one of the classes (Alpaydin, 2009).

Machine learning techniques are being applied to large clinical data tasks (Rumshisky et al., 2016) to create diagnostic models (Hannun et al., 2019) or prediction models (Bayati et al., 2014). The advantage of using machine learning for Stroke outcome predictive model building is the adjustability of models to new data as it becomes available. This property helps to make improvement in the model prediction accuracy.

In recent years, some of the machine learning techniques have been successfully applied in acute stroke management. A study proposed a method for developing a stroke severity index (SSI) by using administrative data. Stroke severity was measured using the National Institutes of Health Stroke Scale (NIHSS). k-nearest neighbours and conventional multiple linear regression (MLR) were used to develop prediction models, the models' performances were comparable according to the Pearson correlation coefficient between the SSI and the NIHSS (Sung et al., 2015). Van Os and colleagues compared multiple machine learning algorithms for the prediction of mRS outcome after endovascular treatment for stroke, with no particular advantage of more modern approaches in comparison with logistic regression, but high prediction performance (van Os et al., 2018). Several works have used machine learning techniques to predict the final infarct voxel-by-voxel, e.g. (Livne et al., 2018).

A biomarker is a measurable indicator of some biological state or condition. Biomarkers may be used alone or in combination to assess the health or disease state of an individual. Blood-based biomarkers are tools to help in the diagnosis of stroke aetiology. Finding stroke biomarker can contribute to improvement in diagnostics (Muñoz et al., 2018).

A study by Sale et al used common inflammatory biomarkers as stroke prognostic factors and designed a model which predicts functional cognitive improvement after rehabilitation treatment from the early stage of stroke. Model inputs included functional data, clinical data, biochemical parameters, and data regarding the health status of rehabilitative treatment, to predict the output of the rehabilitation process. The outcome of stroke was successfully assessed using Multilayer perceptron (MLP), ANN, and SVM (Sale et al., 2018).

Bioinformatics

Bioinformatics starts where the amount of biological data become too large to be processed by a human and researchers develop new ways to analyse it with computers (Yang et al., 2017). It includes the computational analysis of high-throughput biological datasets called omic data sets. There are various types of omics datasets such as genomics (the analysis of complete genomes, i.e. complete sets of genes in a certain species), proteomics (the analysis of proteomes, i.e. complete sets of proteins in a cell), transcriptomics (the analysis of gene expression of complete genomes, i.e. the analysis of complete sets of transcripts) (Tseng et al., 2015). Metabolomics is a term used to describe measurements of multiple small molecule metabolites in biological specimens. It provides a snapshot of the physiology (Sidorov et al., 2019).

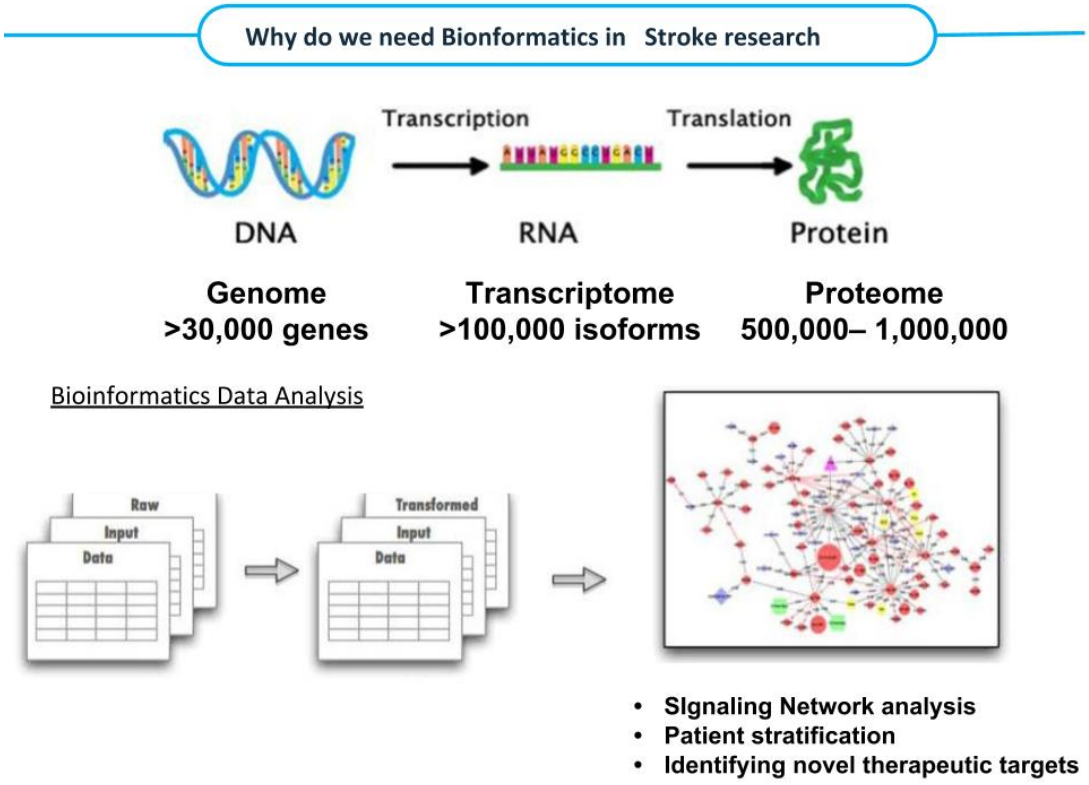

Figure 4: *The process of high throughput biological data for biomarker discovery by creating interaction networks between biologically involved molecules in a disease aetiology*

Using "omics" biomarker discovery approaches (genomics, proteomics, lipidomics or metabolomics), from multicenter prospective study-derived biobanks, offer opportunities for discovery and validation of both novel risk factors and prognostic markers. This opens up a great potential to achieve improved prognostic stratification in stroke (Goldenberg et al., 2014). Studies on metabolome-related stroke discovered several important metabolite-stroke associations (Sidorov et al., 2019).

To illustrate the use of bioinformatics, one can e.g. consider a study by Muñoz et al. They performed a proteomic-related study using a clinical proteomics approach, to identify candidate biomarkers for stroke diagnosis and stroke rehabilitation. This demonstrated a fundamental role of fibrinogen plasmatic levels on patient admission to the stroke rehabilitative care. It was found to correlate with a gain in activities of daily living (ADL) at discharge from rehabilitative care (Muñoz et al., 2018).

With the advances of biological networks (computational representation of biological interactions), it is possible to discover reliable and accurate molecular biomarkers and sub-network biomarkers. In the Muñoz study, they performed proteomic profiling in fresh thrombus samples of the ischemic stroke patients after intra-arterial thrombectomy treatment. Protein-protein interaction networks were generated from proteome expression profiles. These networks approach helped identifying the functional ontology of the highly connected nodes, thrombotic proteins expression correlation to functional modules, and helped with the identification of causal regulators of thrombotic protein modules playing a potential role in thrombus formation (Muñoz et al., 2018).

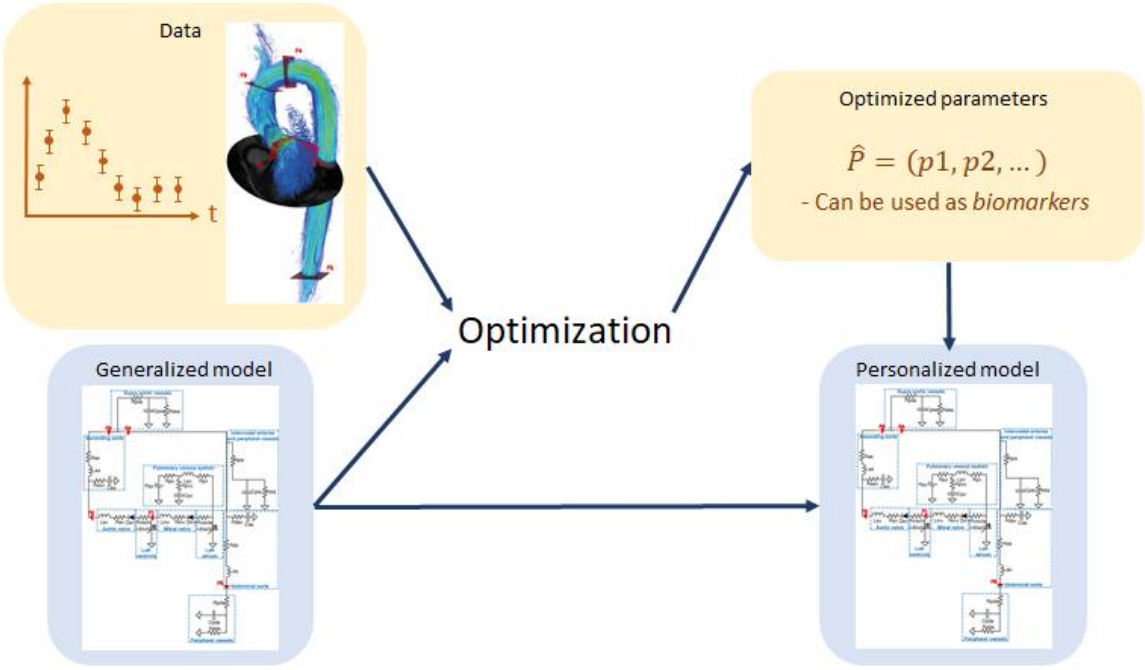

Figure 5: *Overview of the basic principle by which biomarkers can be calculated from data using mechanistic models. The basic principle is essentially the same for all models, but this particular example is taken from modelling of vessels around the aorta, which are calculated using a standard Windkessel model and 4D flow MRI data (Casas et al., 2018, 2017).*

3.2 Multi-level mechanistic models for biomarker calculation and simulation of scenarios

Biomarker calculation using mechanistic models

The basic principle by which biomarkers can be obtained from mechanistic models is depicted in Fig 5. The example is taken from (Casas et al., 2017), but the principle is quite general. The data in this example is taken from a technique called 4D flow MRI, which is based on Magnetic Resonance Imaging (MRI), where the flows can be calculated in all time-points, and where they are both

determined by the direction of the flow, and by the time-point (the flows change quite substantially within each single heartbeat). The various black planes indicate areas in the vessel system where a quantification is made. Each such quantification leads to a time-series, which is symbolically depicted to the left. The time-series is combined with the mechanistic model. The mechanistic model in this example is based on a zero-dimensional ODE model that describes how the blood flow propagates from each compartment in the model to the next, based on various physical properties of the vessels, such as volume, stiffness, compliance, etc. These particular models are based on an analogy between blood flows and electrical circuits and are often referred to as Windkessel models. In any case, by creating a cost function, which quantifies the distance between the experimental data and the simulations, for each parameter combination with values of all such physical properties, the optimal parameters can be obtained by a simple optimization algorithm. In other words, using the data, describing only the flows, individual patient-specific values for physical properties of the vessels can be obtained. Note that these obtained values are not inherently available in the raw data, but that the data needs advanced analysis and lots of knowledge of the physical processes involved for this additional information to be obtained. In this example, the new physical properties of the vessels (compliance, stiffness, volumes, etc) are new biomarkers, which provide additional information about the patient. This particular information says new things about the patient, which are useful, when one e.g. seeks to calculate the risk that the patient will suffer a stroke: the stiffer the vessels, and the smaller the effective volumes, the higher risk for the patient to suffer a cardiovascular event.

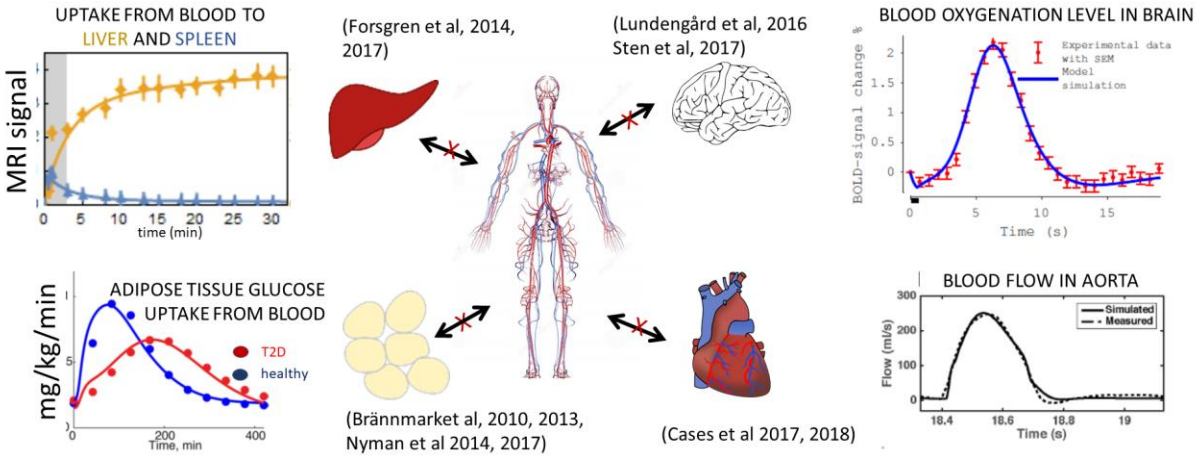

Figure 6: *Overview of some of the most important sub-models that will be included in the multi-level, multi-time-scale, and multi-species models used within the PRECISE4Q project.*

Multi-level, multi-time-scale, mechanistic, and multi-species (M4) modelling to simulate different scenarios

As is indicated in Fig 6, mechanistic models can be combined with each other. In the figure, you see a particular combination that will be done in the PRECISE4Q project: the blood flow models will be combined with mechanistic models for each of the main organs that either secrete or take up things to/from the blood. For instance, the adipose tissue takes up certain metabolites (such as glucose and fatty acids) during postprandial conditions following a meal, and give back certain metabolites (fatty acids but not glucose) during fasting conditions long after a meal is over. Similar uptake and release properties are present for e.g. the heart, the liver, the brain, and the pancreas. In previous projects, we have developed models for each of these organs individually, and they are now being combined into a single interconnected model. An example of what such a model can do is seen in Fig 7. On the long-term time-scale (months), the model can describe slow processes like weight-loss. On the

shorter time-scale (hours), the same model can describe how a meal is digested, by e.g. looking at the time-varying glucose uptake in one of the organs. On the shortest time-scale, one can look at how intracellular processes, such as protein signalling phosphorylation events, happen within seconds or minutes after a hormone has reached a target cell. All of these processes can be described by the same model, and as can be seen in Fig 7, the model (line) can describe the data (dots) with a good quality, both for healthy conditions (blue) and type 2 diabetes (T2D, red). We also have a corresponding model for mice, and a scaling between them, meaning that the resulting model is both multi-time-scale, mechanistic, multi-level, and multi-species (M4).

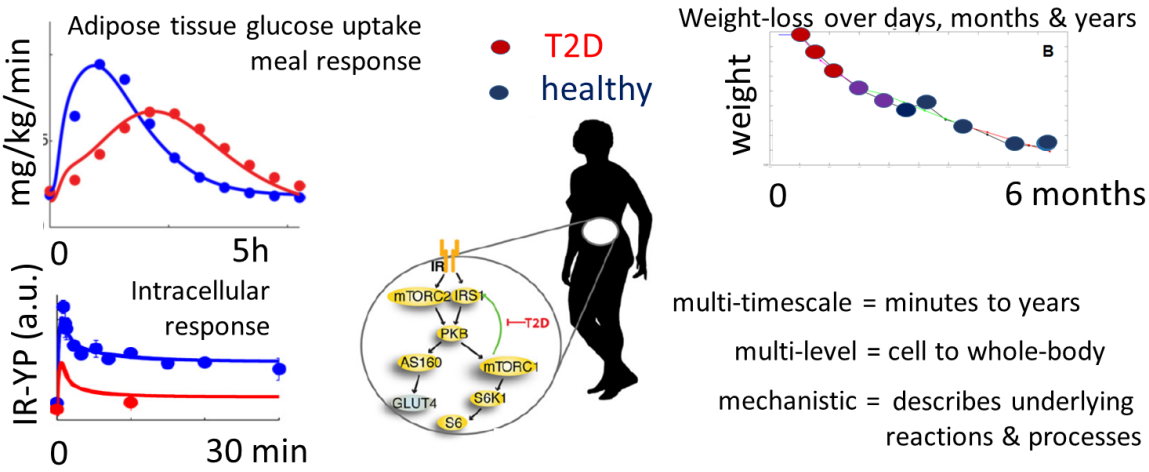

Figure 7: *Overview of the capabilities of our multi-time-scale, multi-level and mechanistic models, to be used within the PRECISE4Q project.*

In addition to these already mentioned models, perhaps most relevant for modelling to help prevent stroke, there are also models for prediction of clinical outcome, and models that are more useful for modelling in the acute treatment phase. CUB has developed a biophysiological model of brain perfusion that integrates individual patient-specific imaging data and boundary conditions (e.g. blood pressure, intracranial pressure, etc.). This model computes flow volumes, velocities, and perfusion pressures for different brain areas. This is the first brain circulation model that integrates individual stroke patient neuroimaging data and enables simulation of brain blood flow and perfusion based on vascular pathology such as stenosis and occlusion. The model has the advantage that it can be adapted to individual patient data derived from medical neuroimaging sources. The model also allows incorporating vessel segments into the model/planar graph to examine their effect in the cerebral vasculature and/or to supplement missing medical data/information. Taken together, the tool provides a dynamic and individualised simulation model for the brain circulation, integrating individual patient data, and enabling simulation of blood flow and perfusion in stroke.

Importantly, this model is based on routine clinical imaging. Thus, it is able to provide individual data for a precision medicine approach without the necessity of additional time-consuming and even potentially harmful approaches used today in perfusion imaging.

However, while this mechanistic model can incorporate some of the features available in current dataset and predict the time-evolution of key biomarkers in a patient-specific manner, there are still many pieces of data that these mechanistic models cannot make use of. Using fusion methodologies, the outputs of our mechanistic model along with the data not accommodated by our mechanistic models will need to be combined with machine learning approaches to create hybrid stroke models. This will allow precise personalized predictions for stroke treatment and outcome.

3.3 Proposed general schemes for hybrid modelling

The main schemes available for hybrid modelling are outlined in Figure 3. The first option is parallel models (Fig 3A), where each of the ingoing models co-exist with the other, but where there is no cross-talk between them. This can e.g. be done by just putting the different models in the same file, and is in many situations a good enough solution, albeit it is almost not to be considered as an actual hybrid model. The second option is called sequential hybrid modelling (Fig 3B), which is done if the output of one of the two models produces some of the inputs used by the other model. This approach is also called a staged hybrid model, which is the terminology we used in the PRECISE4Q application. Sequential modelling can also be done in a reciprocal manner, where the output to the second model serves as the input to the first model, for each step in the model simulations. Such models are also called iterative hybrid models, which is the name we used in the application. Finally, the last option for hybrid modelling is called blended modelling (Fig 3C). Then the two models have been fully merged with each other, into a combined model, that cannot be separated. In the application, we used the term blended modelling. Let us now consider how machine learning, bioinformatics, and multi-level mechanistic modelling can be combined to obtain the two main tasks needed in the PRECISE4Q project: a) calculation of risk factors, b) simulation of scenarios.

4 Our approach for hybrid modelling for calculation of outcomes and risk factors

The approach that we will take for the hybrid modelling, when calculating risk scores and diagnoses is outlined in Fig 8. As can be seen, there are two types of hybrid modelling approaches taken: blended modelling (Step 1), and sequential modelling (Step 2) (Fig 3).

The blended hybrid modelling is done by introducing a statistical and phenomenological component into the mechanistic model. In practice, this will be done by using a modelling methodology called Nonlinear Mixed-Effects Modelling (NLME). More specifically, in NLME models, there is the possibility to introduce covariates into otherwise normal nonlinear ODE models. These covariates impact the value of the parameters in a phenomenological fashion, described by a function. This function can e.g. be something as simple as a sub-division into two distributions, in the case of a Boolean covariate. Say e.g. that some of the parameters in the mechanistic blood flow model (Fig 5) are dependent on the gender of the patient. If that is so, the gender of the patient is the covariate. The data used for estimation of the parameters will then be subdivided into two parts: one for males and one for females. Each gender will then get a specific distribution for those parameters where the covariate is specified as having an effect. A slightly more advanced example is the case where the covariate is continuous. Such a covariate could e.g. be the age of the patient. The age is a property of the patient that does impact the mechanistic parameters (such as the compliance), but we may not yet understand all the mechanisms involved in that dependency. If that is the case, the age should be introduced as a covariate. One will then need to postulate a phenomenological formula for how the covariate is impacting the parameter: choosing between such options as a linear dependency, a saturated dependency (when there is a maximal impact), a step-wise dependency (where there is a sudden jump in a affect at a specific age), etc. This is then a phenomenological component, appearing in an otherwise normal mechanistic model. Finally, NLME models are also more statistical in nature, compared to normal ODE models for another reason: NLME models describe the distribution of each parameter across the population. This distribution is normally described by a covariance matrix, which specifies not only the width of the distribution, but how each parameter is correlated with the other parameters. In practice, this correlation matrix is obtained at the same time as each individual patient, by formulating a joint likelihood function in the estimation step, and

by introducing the additional assumption that the parameter values across the population should follow a certain predefined type (such as normal or log-normal distribution, etc). This additional assumption implies that the parameters will be more well-determined in the case of non-informative data (where the data for an individual patient is insufficient to have well-determined values for all parameters), as demonstrated in e.g. (Karlsson et al., 2015). (Karlsson et al., 2015) also contains a good introduction to the field of NLME, with all the central equations introduced, with practical tips on software, with guidelines on when it is beneficial to use it compared to just using ODE models, and with concrete examples that illustrate the benefits.

The sequential hybrid modelling is the second step of the proposed approach (Fig 8). In this step, all the new biomarkers obtained using either NLME and the mechanistic models (Fig 5, and above), or using the bioinformatics network modelling (Fig 4), are combined with all the other data, containing the rest of the biomarkers. All these biomarkers are then combined into a classical machine learning model, such as a Bayesian Graphical Network, a neural network, etc (see Section 2.1). The machine learning algorithm is trained based on the large clinical study data available in PRECISE4Q (handled in WP2 and WP3), where both all or many of the biomarkers are measured, and where patient outcome is available. In PRECISE4Q, the main outcome of interest is e.g. whether the patient suffered a stroke within the next pre-defined time-period (e.g. 2 years), but the outcome can be changed without any changes to the hybrid architecture used (Fig 8).

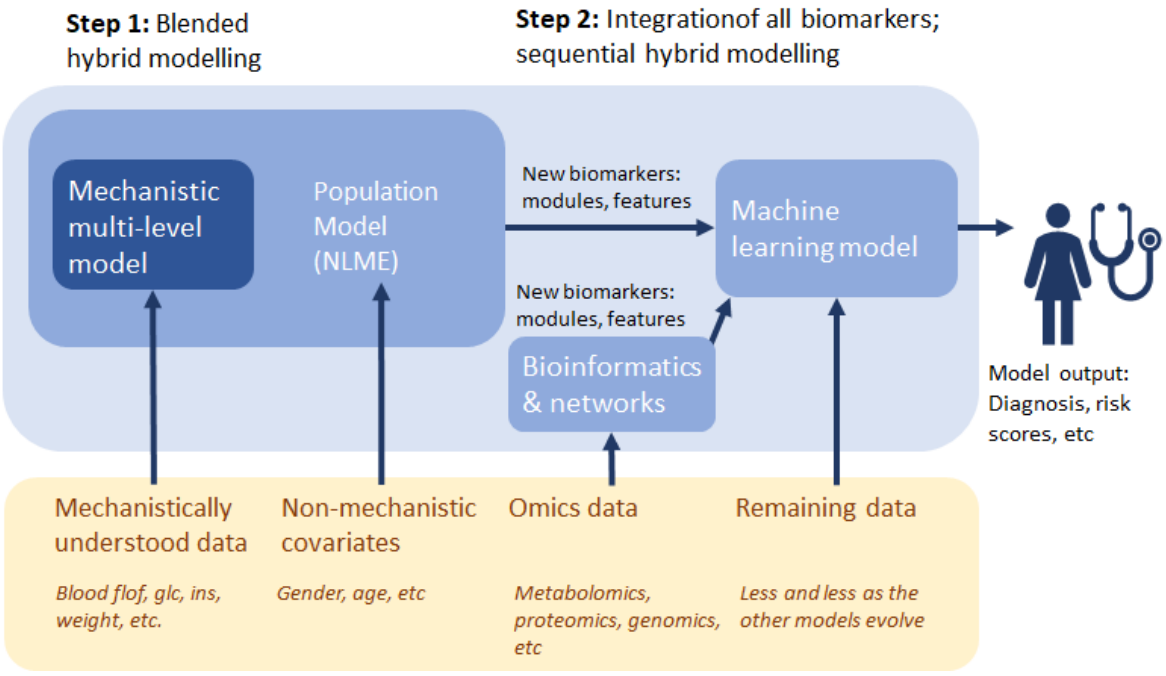

Figure 8: *Outline of our approach for hybrid modelling, when calculating risk scores, diagnosis, etc, of the patient as he/she is at the time of examination.*

5 Our approach for hybrid modelling for simulation of scenarios

The second type of usage for our hybrid modelling is the simulation of scenarios. Our strategy for how these simulations will be done is outlined in Fig 9. As can be seen, there are 8 steps involved.

In Step 1, the data available for the patient that can be entered into the NLME model is identified. Say e.g. that 4D-flow MRI data and some metabolic plasma levels (glucose and triglycerides) are available, that the age and gender are available, but that some other key variables for the model have not been measured (e.g. blood pressure and fatty acid levels). In Fig 9, the available variables are symbolically represented by filled blue circles, and the unavailable variables by non-filled circles.

In Step 2, the non-filled circles are filled in by the usage of imputation or machine learning. In other words, using the large cross-sectional data available in PRECISE4Q, likely values of the non-measured variables can be estimated. Imputation can e.g. be done by finding the most similar patient, and taking his/her values, by using the covariance matrices available in the NLME models, or by simply taking the population average values. More advanced machine learning approaches could develop specific models for how to map the measured variables to the non-measured variables, by considering the measured variables as inputs, and the non-measured variables as outputs (Fig 1B).

In Step 3, the additional values that are needed to simulate the model are added. In Fig 9, these values are symbolized by the filled grey circles. There are different approaches needed for this step. For certain parameters, they can be trained from the available data. This is the case, e.g. for the parameters in the blood flow model, which can be trained from the 4D-flow MRI data (Fig 5). In other cases, parameters have to be taken from either literature values (if they are available as measurements from e.g. a standard human), or one simply needs to estimate a start guess based on experience from previous models. All these approaches imply an uncertainty in the chosen values, meaning that many sets of parameters will be available for the same patient.

In Step 4, all the parameters and values obtained so far are used to initiate the M^4 model. The resulting model is personalized, and is thus sometimes referred to as a digital twin of the patient.

In Step 5, the personalized model is used to investigate different scenarios. Such scenarios could e.g. those depicted in Fig 7, where one can see the predicted response to different drugs and alterations in diet.

In Step 6, these simulations are read out, and the relevant biomarkers and other interesting model properties are analysed. Biomarkers are those that can go into the risk and diagnosis calculations, but there are also many other variables that could be interesting to investigate. Say e.g. that the simulation shows a decrease in inflammation and body fat upon a change in diet. One would then like to know why these positive effects, which will lower the risk of having a stroke, are seen. That question can be answered by inspecting more detailed simulations, that reveal the precise mechanisms and processes that are involved in responding to the diet, and how these mechanisms come into play for this particular patient. Those analyses are interesting and useful, even though they may not be explicit biomarkers, that can be used to calculate the updated risk.

In Step 7, a similar imputation or machine learning scheme as in Step 2 is used to infer values for the other biomarkers needed for the risk engine (Fig 8). These remaining data consist of omics values and other non-mechanistic parameters not given by the model output, and they are symbolically represented by orange and green filled circles respectively.

In Step 8, the full set of data for the chosen time point are used as input into the hybrid risk calculation model (Fig 8), which allows us to infer how e.g. the risk of suffering a stroke has changed in the simulated scenario.

Using these 8 steps, one can thus obtain a personalized model for a patient also in the case of non-complete data, and use the model to both be able to simulate various scenarios, answer the question of why the obtained results are predicted, and calculate the updated risk at each timepoint in the simulated scenarios. This is only possible with the usage of a hybrid modelling scheme, such as the one outlined in Fig 9.

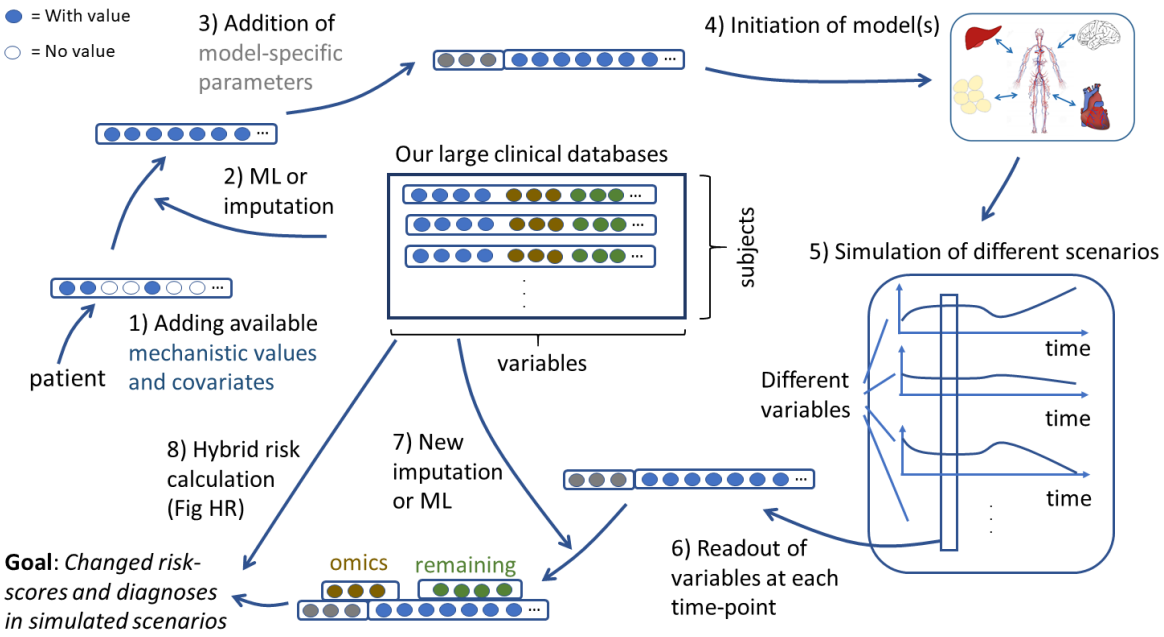

Figure 9: *Outline of our approach to simulation of scenarios, using our new hybrid models.*

6 Summary and outlook

In this review, which constitutes D4.4 in the PRECISE4Q project, we have outlined how machine learning, bioinformatics, and mechanistic modelling will be combined in this project. The background section outlined why hybrid modelling is such a beneficial idea (Fig 2), and what the main available approaches are to such hybrid combinations (Fig 3). In Section 2, we reviewed the three modelling approaches: bioinformatics and machine learning (Fig 4) and multi-level mechanistic models (Fig 5-7). In Section 3 and in Fig 8, we outlined how the hybrid modelling will be done to be able to calculate risk scores and obtain diagnoses. The main idea is to do initial analysis of mechanistically understood data and covariates using advanced NLME approaches (which constitutes a blended hybrid modelling approach), to analyse omics data using bioinformatics and network analysis, which will lead to new biomarkers. These biomarkers will then be used as additional input into the machine learning models, which are the models that in the end calculate the risk scores, such as the risk of suffering a stroke within the next 2 years. Finally, in Section 4 and Fig 9, we outlined an 8-step approach to how we can use our hybrid models to both simulate scenarios, answer the question of *why* the predicted scenarios are believed to be the outcome, and estimate the risk at each timepoint in the simulated scenarios.

There are multiple usages of these hybrid modelling possibilities. The intended usages in PRECISE4Q are outlined in D4.1 and D4.2 where updated use cases have been identified. In the next couple of steps of this project, we will make use of the more and more synthesized databases from WP3, to train and develop our mathematical hybrid models (Step 1-2 in Fig 2). Then, in WP5, we will test and evaluate the performance of our new models on data coming from new clinical studies (Step 3, in Fig 2).

7 References

Alpaydin, E., 2009. Introduction to machine learning. MIT press.

Barbour, D.L., 2019. Precision medicine and the cursed dimensions. Npj Digit. Med. 2, 4.

Bayati, M., Braverman, M., Gillam, M., Mack, K.M., Ruiz, G., Smith, M.S., Horvitz, E., 2014. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. PLOS ONE 9, e109264. https://doi.org/10.1371/journal.pone.0109264

Brännmark, C., Lövfors, W., Komai, A.M., Axelsson, T., El Hachmane, M.F., Musovic, S., Paul, A., Nyman, E., Olofsson, C.S., 2017. Mathematical modeling of white adipocyte exocytosis predicts adiponectin secretion and quantifies the rates of vesicle exo- and endocytosis. J. Biol. Chem. 292, 20032–20043

Brännmark, C., Nyman, E., Fagerholm, S., Bergenholm, L., Ekstrand, E.-M., Cedersund, G., Strålfors, P., 2013. Insulin Signaling in Type 2 Diabetes. J. Biol. Chem. 288, 9867–9880.

Casas, B., Lantz, J., Viola, F., Cedersund, G., Bolger, A.F., Carlhäll, C.-J., Karlsson, M., Ebbers, T., 2017. Bridging the gap between measurements and modelling: a cardiovascular functional avatar. Sci. Rep. 7, 6214

Casas, B., Viola, F., Cedersund, G., Bolger, A.F., Karlsson, M., Carlhäll, C.-J., Ebbers, T., 2018. Non-invasive Assessment of Systolic and Diastolic Cardiac Function During Rest and Stress Conditions Using an Integrated Image-Modeling Approach. Front. Physiol. 9.

Digital Patient Roadmap, n.d. 136.

Doyle, O.M., Tsaneva-Atansaova, K., Harte, J., Tiffin, P.A., Tino, P., Díaz-Zuccarini, V., 2013. Bridging Paradigms: Hybrid Mechanistic-Discriminative Predictive Models. IEEE Trans. Biomed. Eng.

Dzau, V.J., Ginsburg, G.S., 2016. Realizing the Full Potential of Precision Medicine in Health and Health Care. JAMA 316, 1659–1660.

Goldenberg, N.A., Everett, A.D., Graham, D., Bernard, T.J., Nowak-Göttl, U., 2014. Proteomic and other mass spectrometry based “omics” biomarker discovery and validation in pediatric venous thromboembolism and arterial ischemic stroke: Current state, unmet needs, and future directions. PROTEOMICS – Clin. Appl. 8, 828–836.

Gustafsson, M., Nestor, C.E., Zhang, H., Barabási, A.-L., Baranzini, S., Brunak, S., Chung, K.F., Federoff, H.J., Gavin, A.-C., Meehan, R.R., Picotti, P., Pujana, M.À., Rajewsky, N., Smith, K.G., Sterk, P.J., Villoslada, P., Benson, M., 2014. Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome Med. 6.

Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Huang, S., Brooks, M., Lee, M.J., Asadi, H., 2018. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. Am. J. Roentgenol. 212, 38–43.

Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y., 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat. Med. 25, 65.

Hinman, J.D., Rost, N.S., Leung, T.W., Montaner, J., Muir, K.W., Brown, S., Arenillas, J.F., Feldmann, E., Liebeskind, D.S., 2017. Principles of precision medicine in stroke. J Neurol Neurosurg Psychiatry 88, 54–61.

Kelleher, J.D., Mac Namee, B., D’arcy, A., 2015. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

Jonsson, E.N., Karlsson, M.O., Wade, J.R., 2000. Nonlinearity detection: Advantages of nonlinear mixed-effects modeling. AAPS PharmSci 2, 114–123.

Karlsson, M., Janzén, D.L.I., Durrieu, L., Colman-Lerner, A., Kjellsson, M.C., Cedersund, G., 2015. Nonlinear mixed-effects modelling for single cell estimation: when, why, and how to use it. BMC Syst. Biol. 9, 52.

Kim, M., Tagkopoulos, I., 2018. Data integration and predictive modeling methods for multi-omics

datasets. Mol. Omics 14, 8–25.
Lip, G.Y.H., Nieuwlaat, R., Pisters, R., Lane, D.A., Crijns, H.J.G.M., 2010. Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor-Based Approach. Chest 137, 263–272.
Livne, M., Boldsen, J.K., Mikkelsen, I.K., Fiebach, J.B., Sobesky, J., Mouridsen, K., 2018. Boosted Tree Model Reforms Multimodal Magnetic Resonance Imaging Infarct Prediction in Acute Stroke. Stroke 49, 912–918.
Man, C.D., Rizza, R.A., Cobelli, C., 2007. Meal Simulation Model of the Glucose-Insulin System. IEEE Trans. Biomed. Eng. 54, 1740–1749.
Medicine, T.L.R., 2018. Opening the black box of machine learning. Lancet Respir. Med. 6, 801.
Muñoz, R., Santamaría, E., Rubio, I., Ausín, K., Ostolaza, A., Labarga, A., Roldán, M., Zandio, B., Mayor, S., Bermejo, R., Mendigaña, M., Herrera, M., Aymerich, N., Olier, J., Gállego, J., Mendioroz, M., Fernández-Irigoyen, J., 2018. Mass Spectrometry-Based Proteomic Profiling of Thrombotic Material Obtained by Endovascular Thrombectomy in Patients with Ischemic Stroke. Int. J. Mol. Sci. 19, 498.
Nyman, E., Brännmark, C., Palmér, R., Brugård, J., Nyström, F.H., Strålfors, P., Cedersund, G., 2011. A Hierarchical Whole-body Modeling Approach Elucidates the Link between in Vitro Insulin Signaling and in Vivo Glucose Homeostasis. J. Biol. Chem. 286, 26028–26041.
Nyman, E., Rajan, M.R., Fagerholm, S., Brännmark, C., Cedersund, G., Strålfors, P., 2014. A Single Mechanism Can Explain Network-wide Insulin Resistance in Adipocytes from Obese Patients with Type 2 Diabetes. J. Biol. Chem. 289, 33215–33230.
Rappoport, N., Shamir, R., 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. bioRxiv.
Rostanski, S.K., Marshall, R.S., 2016. Precision Medicine for Ischemic Stroke. JAMA Neurol. 73, 773–774.
Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V.M., McCoy, T.H., Perlis, R.H., 2016. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Transl. Psychiatry 6, e921.
Sale, P., Ferriero, G., Ciabattoni, L., Cortese, A.M., Ferracuti, F., Romeo, L., Piccione, F., Masiero, S., 2018. Predicting Motor and Cognitive Improvement Through Machine Learning Algorithm in Human Subject that Underwent a Rehabilitation Treatment in the Early Stage of Stroke. J. Stroke Cerebrovasc. Dis. 27, 2962–2972.
Sidorov, E., Sanghera, D.K., Vanamala, J.K.P., 2019. Biomarker for Ischemic Stroke Using Metabolome: A Clinician Perspective. J. Stroke 21, 31–41.
Stéphanou, A., Volpert, V., 2016. Hybrid Modelling in Biology: a Classification Review. Math. Model. Nat. Phenom. 11, 37–48. https://doi.org/10.1051/mmnp/201611103
Sung, S.-F., Hsieh, C.-Y., Kao Yang, Y.-H., Lin, H.-J., Chen, C.-H., Chen, Y.-W., Hu, Y.-H., 2015. Developing a stroke severity index based on administrative data was feasible using data mining techniques. J. Clin. Epidemiol. 68, 1292–1300.
Tseng, G., Ghosh, D., Zhou, X.., 2015. Integrating omics data. Cambridge University Press.
van Os, H.J.A., Ramos, L.A., Hilbert, A., van Leeuwen, M., van Walderveen, M.A.A., Kruyt, N.D., Dippel, D.W.J., Steyerberg, E.W., van der Schaaf, I.C., Lingsma, H.F., Schonewille, W.J., Majoie, C.B.L.M., Olabarriaga, S.D., Zwinderman, K.H., Venema, E., Marquering, H.A., Wermer, M.J.H., M.C.R., 2018. Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms. Front. Neurol. 9.
VoosenJul. 6, P., 2017, Pm, 2:00, 2017. How AI detectives are cracking open the black box of deep learning. Sci. AAAS.
Warwick, K., 2013. Artificial intelligence: the basics. Routledge.
Warwick, K., 2004. March of the machines: the breakthrough in artificial intelligence. University of Illinois Press.

Wasserman, P.., 1993. Advanced methods in neural computing. John Wiley & Sons, Inc.
Yang, A., Troup, M., Ho, J.W.K., 2017. Scalability and Validation of Big Data Bioinformatics Software. Comput. Struct. Biotechnol. J. 15, 379–386.