PRECISE4Q

PREDICTIVE MODELLING IN STROKE

Data Schema Designs for Each Model

Project Acronym: **Precise4Q**

Grant Agreement number: **777107**

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

D4.3– Data Schema Designs for Each Model

Revision: 1.0

Authors and Contributors	Authors: John D. Kelleher (TU Dublin), Attia Fatima (TU Dublin) Contributor (Catalina Martinez Costa, Julia Amann, Alejandro Garcia, Gunnar Cedersund, Tilda Herrgårdh, Vince Madai)		
Responsible Author	Prof. John D. Kelleher	**Email**	john.d.kelleher@tudublin.ie
Beneficiary	TU Dublin	**Phone**	+353 1 4024789

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
1	06/08/19	Attia Fatima John Kelleher	TUD	Content list
2	20/08/19	Attia Fatima John Kelleher	TUD	Completed First draft
3	05/09/19	Attia Fatima John Kelleher	TUD	Completed Second draft
4	13/09/19	Attia Fatima John Kelleher	TUD	Final Draft Submitted for Internal Review
5	17/09/19	Julia Amann	ETH	Detailed Review and Feedback
6	17/09/19	Catalina Martinez Costa	MUG	Detailed Review and Feedback
7	18/09/19	John Kelleher	TUD	Reworking Executive summary
8	19/09/19	John Kelleher	TUD	Reworking Introduction
9	23/09/19	John Kelleher	TUD	Reworking and Extending Stroke Prevention Section
10	24/09/19	John Kelleher	TUD	Reworking and Extending Acute Model D4.6 Section
11	25/09/19	John Kelleher	TUD	Reworking and Extending Acute Model D4.7 Section
12	25/09/19	Catalina Martinez Costa	MUG	Reviewing Data Tables
13	26/09/19	John Kelleher	TUD	Reworking and Extending Rehabilitation Model Section
14	26/09/19	Catalina Martinez Costa	MUG	Proofing of Entire Document
15	27/09/19	John Kelleher	TUD	Final Editing

Date of delivery	Contractual:	30.09.2019	Actual:	30.19.2019
Status	final **X** /draft ☐			

Abstract (for dissemination)	This document presents the deliverable 4.3 of work package 4 data schema design mainly involving input features selection for stroke prediction models
Keywords	Risk factor, model input features, SNOMED CT, Stroke Summary tables

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 Contents

Executive Summary 6

1 Introduction 7

2 Stroke Prevention (Data Schema for Model D4.5) 10

2.1 Demographic Factors 10

2.2 Obesity 11

2.3 Alcohol intake/Smoker 11

2.4 Physical Exercise 11

2.5 Socioeconomic Status (SES) 11

2.6 Diet 11

2.7 Blood pressure / Hypertension 12

2.8 Stress 12

2.9 Depression 12

2.10 Sleep Disorders and Patterns 13

2.11 Diabetes 13

2.12 Atrial fibrillation 13

2.13 Carotid Artery Disease 13

2.14 (Cardio)-vascular conditions 14

2.15 Dyslipidemia 14

2.16 Thyroid function 14

2.17 Renal function 15

2.18 Inflammation and infection 15

2.19 Genetic Factors 15

2.20 Table 1.1 Data Schema for Prevention Model 16

3 Acute Stroke Treatment Models 19

3.1 D4.6 Hybrid model predicting short term stroke outcome 20

3.1.1 Neuroimaging 20

3.1.2 Demographic Data and Obesity 21

3.1.3 Time Interval Since Onset of Symptoms 22

3.1.4 Stroke Type Diagnosis 22

3.1.5 Stroke Level Severity 22

3.1.6 Treatment 23
3.1.7 Table 2.1 Data Schema for Acute Stroke Treatment Models 23
3.2 Personalised Acute Stroke Quality of Life Prediction Model (D4.7) 24
3.2.1 Neuroimaging 24
3.2.2 Demographic Data 24
3.2.3 Time Interval Since Onset of Symptoms 25
3.2.4 Stroke Type Diagnosis 25
3.2.5 Stroke Level Severity 25
3.2.6 ADL/Accommodation Before Onset 25
3.2.7 Associated Medical Condition (Risk Factors) 25
3.2.8 Swallowing Function/Speech Examination 25
3.2.9 Pharmaceutical Treatment 26
3.2.10 Surgical Treatments 26
3.2.11 Data Schema for Acute Stroke Treatment Models 26
4. Personalised rehabilitation model (D4.8) Data Schema 29
4.1 Functional Assessment 30
4.2 Cognitive Assessment 30
4.3 Cognitive Rehabilitation Training 31
4.4 Fragility 32
4.5 Data Schema for Rehabilitation Model 33
5 Conclusions 35
References 36

List of Figures

Figure 1 Illustration of the decomposition of a modelling task into a set of features 8
Figure 2 Schematic of the target architecture for model D4.6 20

Executive Summary

The primary focus of Task 4.3 was defining the relevant model input features. A foundational concept for this task, and hence this deliverable as a whole, is the distinction between a *factor* and a *feature*. Whereas a *factor* is a concept that a domain expert may identify as relevant for a given prediction scenario, a *feature* is a categorical (nominal or ordinal) or continuous (integer) unit of measurement that may be directly used to train and validate computational models. In other words, a factor is operationalized by one or more features. For example, although *Hypertension* may be identified as an important *factor* to consider in making a particular prediction, for the purposes of developing a predictive computational model, high-level domain factors must be represented by one or more concrete features (i.e. continuous or categorical model inputs). Examples of features that may be used to represent the *hypertension* factor within model inputs include: Hypertension diagnosis (i.e. dichotomous categorical outcome, Yes/No) and *peripheral pressure (i.e. continuous (integer) outcome, mmHG (millimeters of mercury above surrounding atmospheric pressure))*.

This activity builds upon, and extends, previous Precise4Q deliverables that have identified the model use cases and relevant factors. In particular the following deliverables informed this work:

- D1.1 SoA for stroke risk factors prognosis and outcomes
- D1.3 Use cases and their inputs-outputs specification
- D2.1 Overview of data sources and a plan to access available data sources
- D4.1 White paper on stroke risk, health and resilience factors
- D4.2 QOL targets for models created in T4.5, T4.6, T4.7 and T4.8

This deliverable was also informed by workshops and engagements with domain knowledge experts. This work-package extends upon earlier deliverables by mapping previously identified factors of relevance to appropriate feature sets, with model-specific sets representing the initial data schema for the model training phase. Development of data schemas was undertaken by combining information from defined use cases based on patient scenarios (D1.3), with the set of factors identified as relevant to a stroke patient profile extracted for each individual use case. Note that in many instances there is a 1-to-many relationship between factors and features. Accordingly, the current deliverable is summarised as an exercise in appropriately identifying and mapping use-cases, to factors, to features. Input features identified within this deliverable will assist in developing a data-driven understanding of stroke phases and production of predictive models which may be used to facilitate robust case-specific decision support systems for all stakeholders (e.g. clinicians to social workers) involved at the various stages of a patient's journey.

1 Introduction

The overarching objective of PRECISE4Q is the development of a set of data-driven predictive models which specifically target the four stages along a stroke patient's trajectory (*prevention*, *stroke-treatment*, *stroke- rehabilitation*, *stroke-reintegration*). The key to any successful data-driven modelling project is knowing *what to measure* and *how to measure it*. Consequently, feature selection represents a fundamental process in model development. Feature selection is particularly important within a medical context due to the sheer volume and range of clinical assessments used to evaluate patient condition and assess likely long-term outcomes. For example, evaluation of patients during the acute stroke phase is frequently undertaken using imaging data (e.g. neurophysiological and neuroimaging biomarkers) which correlate changes in brain structure and perfusion patterns to assist in predicting clinical outcomes and/or recovery. Conversely, rehab phase evaluations typically employ differing assessments, such as (1) motor ability (e.g. Fugl-Meyer); (2) functional performance (e.g. Wolf Motor Function Test) or (3) self-reported motor activity (e.g. Motor Activity Log, Functional Independence Measure). Accordingly, the development of reliable predictive models is predicated on identification of the most relevant features to use as model inputs.

In the PRECISE4Q project, feature selection has been framed as a process of iteratively decomposing high-level domain *factors* into concrete features. Factors identified from the literature and through engagement with domain experts have been delineated into three primary categories, namely risk factors, health factors and resilience factors. Risk factors for primary stroke are divided into modifiable and non-modifiable risk factors. Both are critical for model development, with modifiable risk factors considered particularly significant as they permit development of preventative interventions. An important advantage of framing feature selection as a factor decomposition process is that it comprises cross-disciplinary dialogue between clinicians, data harmonization experts, and computational modelers. Figure 1 below illustrates the decomposition of a modelling challenge into a set of relevant *factors* (Domain Subfactors) and then onto concrete *features*. For example, *Hypertension* may be identified as an important *factor* for a given modelling task, with Hypertension potentially represented by a number of features within the inputs to the model, such as: *peripheral pressure*, *LVEDP blood pressure over 24h*, *diastolic blood pressure during the day*, and so on.

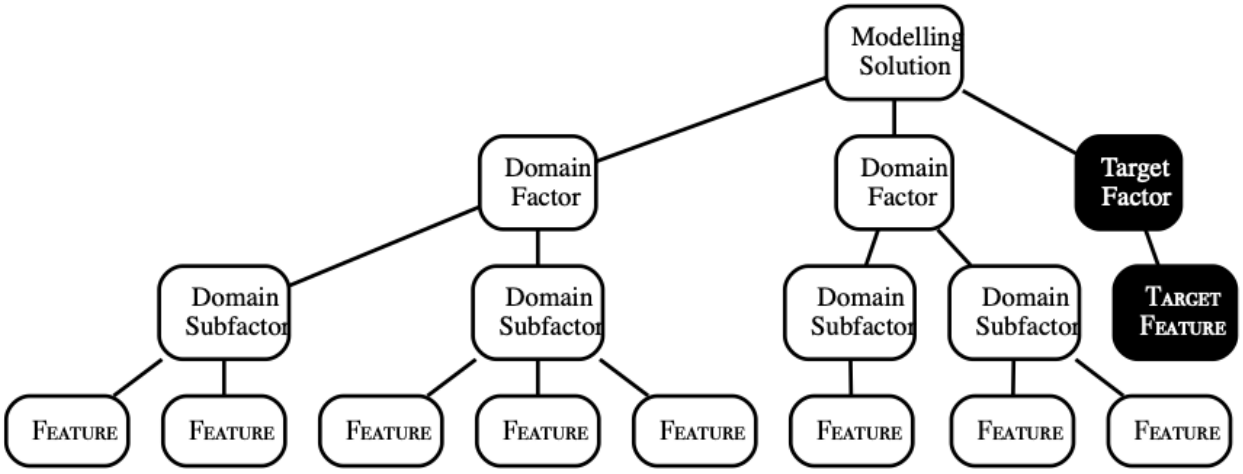

Figure 1 Illustration of the decomposition of a modelling task into a set of features

The point of departure for the PRECISE4Q feature selection process was the consolidation of the factors identified in *D1.1 SoA for stroke risk factors prognosis and outcomes*, *D1.3 Use cases and their inputs-outputs specification*, and *D4.1 White paper on stroke risk, health and resilience factors.* These factors include the aforementioned health, risk, and resilience factors, in addition to life events in stroke affecting well-being (integrated "quality-of-life-concept"). Selection of these factors also involved literature reviews from current state-of-the-art knowledge about stroke risk factors (incl. genetics), prognosis and outcomes after interventions and domain experts' suggestions, i.e. iterations of dialogue between clinicians, care providers and modelers (Delphi Study). This consolidation process resulted in the identification of relevant high-level factors for each of the four stages of stroke treatment.

The current deliverable identifies feature sets which represent each of the factors identified as being relevant for each model. The PRECISE4Q models will be built using the harmonized datasets created in WP3. This data is being aggregated and integrated from various sources within the project. Consequently, the survey of PRECISE4Q data-sources reported in *D2.1 Overview of data sources and a plan to access available data sources* in addition to the SCT- (SNOMED Clinical Terminology) based harmonized stroke summary data prepared during WP3 D3.1 which were used to identify candidate features. Stroke summary data tables (WP3) were used to retrieve input features for each of the 4 stroke phase models corresponding to the factors from the harmonized data.

The current report presents the developed data schemas (feature sets) for the models being created for each of the four phases of stroke, and as such, is structured as follows:

1. D4.5 Personalised Stroke Prevention Model Data

2. D4.6 Hybrid Model Predicting Short-Term Stroke Outcome and D4.7 Hybrid Model Predicting Post-treatment Quality of Life Stroke Outcome Data Scheme
3. D4.8 Personalised Rehabilitation Model Data Schema
4. D4.9 Model Predicting Long-Term Reintegration and Well-Being Data Schema

Each section provides an (i) overview of the *factors* identified as being relevant to the associated model, and (ii) the list of *features* that the factor has been mapped to. Subsequently, each section concludes with a tabular presentation of proposed data schema for each model. Each row in the data schema specified the high-level factor that a feature is associated with, the feature name and the feature type (we have distinguished by nominal, ordinal, boolean, numeric and DateTime). Note, in order to avoid overly long tables, if multiple features share the same factor and type, they are bundled and listed in the same row. The final column in each table may include explanatory notes for the feature, or list the values that the feature can take; in some instances the feature values are mapped to the relevant SNOMED CT (Systematized Nomenclature of Medicine -- Clinical Terms) concept code (Spackman et al., 1997). Ordinal types correspond to named values in which the order is important (e.g. values for the FIM (Functional Independence Measure) scale). Note, as we describe in the next paragraph, the data sources for each model vary. One result of this is that the feature definitions can vary across data schemas because the feature definition is tied to the data source it is taken from.

A key consideration in the development of an integrated data frame for any modelling project is data availability. Accordingly, in order to ensure that the presented data schema is both medically effective and possible to collate (i.e. available) the feature set for each model is drawn from a primary data source. For example, for the stroke prevention model the Cardiovascular Risk Factors in Patients with Diabetes - a Prospective Study in Primary Care (CARDIPP) dataset has been used to guide variable selection (ClinicalTrials.gov Identifier: NCT01049737); the data selection for the acute stroke treatment model D4.6 is primarily drawn from the data available at Charité Universitätsmedizin Berlin, whereas the dataset for the second acute stroke treatment model (D4.7) is drawn from the Swedish Stroke Registry (Riksstroke); and the feature sets for the rehabilitation and reintegration models are primarily based on the data available from Institut Guttmann, Hospital de Neurorehabilitación. The data integration and harmonisation process is in progress and more data sources will be integrated, expanding the feature sets for each model. For example, it is anticipated that genomic data will be integrated into several models using data from the Estonian Genome Center at the University of Tartu. In addition, for the prevention and reintegration models we are exploring the integration of longitudinal medical history data from the Swedish Stroke Registry (Riksstroke) and insurance history data from the Allgemeine Ortskrankenkasse Nordost (AOK).

2 Stroke Prevention (Data Schema for Model D4.5)

Model D4.5 is designed to support patient screening outside of a hospital setting. The primary prediction target for model D4.5 is: *an individualised risk of stroke with a parameterised prospective time-period of 3 to 5 years*. This model addresses both primary (prevention of stroke in patients who have not previously suffered from stroke) and secondary (prevention of recurrence of stroke) prevention. Due to the marked difference associated with the risk factors, health factors and resilience factors of primary and secondary stroke patients, both use cases will require individual prevention model development. For example, the secondary prevention model will utilize the additional patient-specific data available post-treatment (e.g., neuroimaging, etc.) and the inherently increased risk factors, when making a prediction. For the purposes of the current (pre-modelling) data schema, the initial input features are defined, which focus on primary prevention. As modelling work progresses, and additional data frames are identified and made available, the current set of input features will increase.

2.1 Demographic Factors

A number of demographic features have been found to be discriminative in terms of stroke. For example, **age** represents the most significant single determinant of stroke, with the risk of stroke doubling every decade above 55 years of age (Boehme et al., 2017; Johnston et al., 2009). However recent research reports increase in stroke cases in young individuals with this trend projected to continue through to 2025 (Béjot et al., 2016). This factor is thus represented in the model inputs with the **Age** feature and recorded as an ordinalised variable of chronological age in 10-year (decile) ranges. **Sex** is an unmodifiable risk factor and has significant relevance in the pathophysiology and phenotype of stroke. Generally males exhibit a higher risk of stroke in early and middle adulthood, with a shifting prevalence among women among the elderly sub-population (Benjamin et al., 2019). This can be linked to age where a rising proportion of females suffer strokes above 85 years of age. Also stroke related mortality is higher in older females (Haast et al., 2012). In younger women pregnancy and the post-partum state are also associated with stroke risk (Boehme et al., 2017). This factor is represented in the model inputs by the **Sex** feature (see Table 1.1 for details). **Ethnicity** has been correlated with stroke occurrence. Studies have reported a higher risk for stroke for African Americans, Latino Americans, and American Indians compared to other ethnicities. A potentially confounding factor here is that this correlation might be biased by socio economic factors as well (Boehme et al., 2017; García et al., 2017). This factor is represented in the model inputs by the **Racial Group** feature which is a nominal feature which categorizes a subject into 1 of 80 possible values for a subject. We also include the **Marital Status** of the patients.

2.2 Obesity

Obesity is a risk factor of stroke. However, obesity is a composite parameter with associations to other risk factors. **Body-mass-index (BMI)** is a commonly used parameter to measure obesity. We note that the disadvantages of the BMI as a parameter have been increasingly discussed. For example, some discussions suggest the waist-to-hip-ratio as being a better parameter to predict stroke risk. However, notwithstanding potential issues with the BMI measure, it is still commonly used as a feature to judge obesity and is included in the data schema to reflect this factor; partly because its wide usage makes the feature available in many datasets.

2.3 Alcohol intake/Smoker

Studies have shown an association between alcohol and other substance abuse and stroke. Sometimes this association may be indirect (collinear/causative with other features), for example heavy drinking is also associated with hypertension which in turn can lead to stroke. We capture this factor in the model feature set with the ***Alcohol intake (How often)*** *which encodes an ordinal categorization of the subjects' alcohol intake.* Cigarette smoking falls under a modifiable risk factor that is independently associated with ischemic stroke. Studies have shown a direct correlation between dose over time and the risk of stroke (Wilke et al., 2015). This factor is by recording the number of years a person has been smoking (**Smoking Years**).

2.4 Physical Exercise

The protective effect of regular physical exercise on decreasing stroke risk has been well established. Several studies have reported high correlations between levels of physical exercise and decreased stroke risk (Boehme et al., 2017; O'Donnell et al., 2010). The CARDIPP dataset records a number of ordinal features capturing different aspects of this factor (see Table 1.1 for details): ***Is your daily work physically straining***, ***Exercise in the last 12 months***, ***Time for exercise that makes you warm***

2.5 Socioeconomic Status (SES)

It has been reported that a lower personal and community/locale socioeconomic status (SES) may significantly increase risk for adverse clinical outcomes among ischemic stroke patients (Yan et al., 2017). We capture socioeconomic status (SES) through the **employment**, **returning to work**, and **education** features.

2.6 Diet

Diet can directly influence vascular risk factors. Diets high in saturated fat, trans fat and cholesterol can raise blood cholesterol levels. Diets high in sodium (salt) can increase blood pressure and diets with high calories can lead to obesity. The INTERSTROKE study found a

reduction in stroke risk associated with fruit and fish consumption and increased stroke risk with foods like red meat, organ meat, eggs, fried foods and salty snacks (O'Donnell et al., 2010). A large 17 year cohort study, comprising 395,048 persons, found that good diet is associated with a significant reduction stork risk (Paterson et al., 2018). Model input feature for diet is represented by an ordinal factor with the following attribute values: ***High Fat Diet, Low Fat Diet, Healthy Diet***

2.7 Blood pressure / Hypertension

Hypertension (modifiable by life-style and medication), leads to an increased risk of stroke and the incidence of hypertension increases with age. Thus, treatment of hypertension is an effective measure to reduce stroke risk. However, hypertension treatment is still far from optimal in developed countries and low-income countries have the highest prevalence of elevated blood pressure (http://www.who.int/features/qa/82/en/). Recent strategies to assess this biomarker based on the variability in blood pressure measurements over time have been shown to be a better predictor of risk than the static snap-shot measurements (Rothwell et al., 2011). Consequently, we include a number of different features associated with hypertension within the data schema, including measurements over time. The following features from the CARDIPP dataset are associated this factor: ***Hypertension, Peripheral pressure, Central peripheral Pressure, Peripheral pressure ratio, Pulse wave velocity (PWA), Left ventricular end diastolic pressure (LVEDP), LVEDP Systolic blood pressure over 24h, LVEDP blood pressure over 24h, Systolic blood pressure during day, Diastolic blood pressure during day, Systolic blood pressure during night, Diastolic blood pressure during night, Radial mean arterial pressure***

2.8 Stress

The presence of psychosocial stress, defined as an imbalance between demands placed on us and our ability to manage them, has been connected with unhealthy lifestyles. Furthermore, its associated stress hormone release biologically affects the body and is reported as a partially modifiable risk factor for stroke (Everson-Rose et al., 2014). Consequently, it is important to include features for stress in the data schema for future modelling work. The CARDIPP dataset records a number of features related to stress (including: *Stressed, Spirited and strong, Very nervous, Calm and harmonious, Full of Energy, Gloomy and sad, Anxiety worries or anxieties*). However, at present we simply use a single nominal feature **Stress** which records whether stress is *present*, *absent* or *unknown*.

2.9 Depression

Depression a modifiable risk factor has been associated with increasing incidence of stroke (Pan et al., 2011). The INTERSTROKE study on 10 stroke risk factors reported depression to be an

important risk factor (O'Donnell et al., 2010). The model input features include an ordinal feature **Depression**.

2.10 Sleep Disorders and Patterns

Observational and theoretical considerations suggest a link between sleep disorders and vascular event risk (Koo et al., 2018). Sleep disorders, including insomnia and sleep-related movement disorders, are highly prevalent in patients at risk for stroke, and obstructive sleep apnea has been linked to increased stroke risk. Also, sleep disorders are linked to increases in the prevalence of stroke risk factors (Phua et al., 2017), which might be at least one pathway through which stroke risk is mediated. This factor is represented in the model features by a nominal feature recording whether the patient is complaining about ***Insomnia*** or not, or whether it is unknown.

2.11 Diabetes

Diabetes is a major risk factor for stroke occurrence with up to 20% of diabetes patients dying of stroke (Boehme et al., 2017). Indeed, it has been suggested that the increase in diabetes prevalence within younger populations might be explanatory for the overall increasing incidence of stroke among this sub-population (Kissela et al., 2005). Model Input features for diabetes from the CARDIPP dataset record the number of years since a patient was diagnosed with diabetes (***Diabetes duration***) and also ***Insulin, fP Glucose,*** and ***Wide range CRP***

2.12 Atrial fibrillation

Atrial fibrillation is one of the largest risk factors for stroke, although this is modifiable by medication and surgery. The explanation of the correlation of atrial fibrillation and stroke based on blood clot generation due to stasis of blood in the left atrium has been challenged, and research is underway to create better models for correlating atrial fibrillation with stroke. However, notwithstanding the challenges in explaining the correlation, research reports a substantial reduction in stroke risk after the treatment of atrial fibrillation. Atrial fibrillation is directly captured using a Boolean feature that is categorized via the **Cardiovascular Condition** factor (see discussion below), however two other model input features associated with atrial fibrillation have been proposed to provide more detail on the atrial fibrillation factor, namely:

- ***Size of left atrium from hear echo examination***
- ***Size of septum from heart echo examination***

2.13 Carotid Artery Disease

Carotid artery disease, or large vessel atherosclerotice, primarily affects stenosis of the internal carotid artery. This is a modifiable stroke risk and medication or surgery such as carotid endarterectomy (CEA) can reduce the risk of stroke. In novel therapies functional images are

used to estimate plaque vulnerability and this has been shown to be a better predictor of stroke, compared with traditional therapy that uses lumen width as determinant of stroke risk. This factor is represented in the model inputs using two Boolean features:

- ***Plaque from vascular examination, dexter (right)***
- ***Plaque from vascular examination, sinister (left)***

2.14 (Cardio)-vascular conditions

It is well established that (cardio-)vascular disease represents a risk-factor for stroke. Peripheral artery disease (PAD) (Banerjee et al., 2010), myocardial infarction and vascular diseases in other body territories increase the risk of stroke. Also heart failure has been associated with stroke risk (Kim and Kim, 2018). This risk factor is partially modifiable by life-style-change. The importance of this factor is reflected in the model inputs by the fact that a relatively large number of features are employed to encode it, including: ***Angina, Atrial fibrillation, Major Cardiovascular Event Status, Heart failure, Heart attack, Stroke, Systolic left ventricular function from echocardiography, LVEDP left ventricular function from Doppler velocity, Left ventricular function from Tissue Velocity Imaging, Echocardiography of pulmonary veins during systole, Echocardiography of pulmonary veins during diastole,*** **Enlarged heart (left ventricle divided by body surface)**

2.15 Dyslipidemia

Dyslipdemia of cholesterol and triglycerides has direct association with stroke occurrences. While higher LDL levels raises the risk of ischemic stroke, low HDL levels can also contribute to stroke risk. Furthermore, the ischemic stroke atherosclerotic subtypes are strongly associated with dyslipidemia. Model input feature for dyslipidemia include circulatory biomarkers including: ***S-Triglycerides, S-Cholesterol, S-HDL Cholesterol, S-LDL Cholesterol, Ratio HDL/LDL, Apolipoprotein A1***

2.16 Thyroid function

Subclinical thyroid dysfunction is a common endocrine condition among the general population, including a prevalence reaching up to 15% for subclinical hypothyroidism (SHypo), and 12% for subclinical hyperthyroidism (SHyper). The cardio-cerebral vascular system is one of the major targets of thyroid hormones. SHypo has been shown to propagate vascular risk factors, such as hyperlipidemia, metabolic syndrome and vascular stiffness. SHyper has been proved to promote vascular damage in numerous ways, including facilitating hypertension, maintaining hypercoagulable state and causing endothelial dysfunction (Zhang et al., 2019). Thyroid function is represented in the model input features through measured ***Parathyroid hormone level*** and ***Vitamin D***.

2.17 Renal function

It has been reported that stroke risk is increased among subjects characterised by lower baseline measurements of renal function (estimated glomerular filtration rate (eGFR) and total proteinuria or albuminuria), when adjusted for variables known to influence stroke risk (Sandsmark et al., 2015). CARDIPP records a number of subject attributes associated with renal function: ***S-Creatinine level, S-Potassium level, Cystatin Filtration rate, Createnin in index mmol/L, Albumin to Createnin Index (in Urine), Albumin in urine, g/l***

2.18 Inflammation and infection

Inflammation and infections have been reported risk factors for ischemic stroke. Proinflammatory alterations cause thrombogenesis through inflammatory stimulation in the cerebral vascular system. In early atherogenesis inflammatory cells accumulate in vascular wall and get activated, later resulting in plaque rupture and thrombus formation leading to stroke. Biomarkers of Inflammatory markers (such as leukocytes, fibrinogen, and C-reactive protein) are good predictors of ischemic stroke. Risk of stroke is higher among those with chronic infections. Acute chronic infection can activate coagulation and contribute to atherogenesis. A number of inflammatory cytokines contribute to these phenomena including **interleukin-10** and **interleukin-6**.

2.19 Genetic Factors

Genetic factors play an important role in stroke risk, as shown by partial heritability (30% risk increase due to family history). However, identification and quantification of these factors is challenging due to the high levels of heterogeneity associated with stroke causes and populations (Boehme et al., 2017). One can distinguish between single gene disorders, where stroke is the primary manifestation of the diseases and genetic variants associated with ischemic stroke. The former include diseases like CADASIL, CARASIL, sickle cell disease, Fabry disease and others (Boehme et al., 2017), for example, several studies have reported associations between stroke and the ABO blood type gene. Other gene loci have also identified in the literature, but here the disease mechanisms are unclear and are a focus of investigation (Boehme et al., 2017). Genetic history is a complex factor to model, thus we have identified two primary features for inclusion in the data schema to represent genetic factors, both of which are currently available within the Swedish CARDIPP data frame, as follows (see Table 1.1 for details):

- ***Angiotensinogen Single nucleotide polymorphism*** (***AGT SNP)***
- ***Renin genotype***

2.20 Table 1.1 Data Schema for Prevention Model

Factor	Feature	Type	Notes, Example Values (SNOMED CT)
Demographic	Age (Patient's age)	Ordinal	Chronological age in 10-year (decile) ranges
Demographic	Sex (Patient's Biological sex)	Nominal	Female (SCT: 248152002) Male (SCT: 248153007)
Demographic	Ethnicity	Nominal	Patient categorization into 1 of 80 Racial Groupings - subclass of (SCT: 415229000)
Demographic	Marital Status	Nominal	Marital status: single, never married (SCT: 125725006) Cohabiting (SCT: 38070000) Divorced (SCT: 20295000) Widowed (SCT: 33553000) Separated (SCT: 13184001)
Obesity	Body Mass Index	Numeric	BMI (SCT: 60621009)
Alcohol intake/ Smoker	Alcohol Intake (How often?)	Ordinal	Finding of alcohol intake (finding) (SCT: 365967005) + Known absent (qualifier value) (SCT: 410516002); Finding of alcohol intake (finding) (SCT: 365967005) + Unknown (qualifier value) (SCT: 261665006); Alcohol intake within recommended daily limit (finding) (SCT: 428202005); Alcohol intake exceeds recommended daily limit (finding) (SCT: 429775004)
Alcohol intake/ Smoker	Smoking Years	Numeric	
Physical exercise	Time for exercise that makes you warm	Ordinal	Distinguishes between 6 levels
Physical exercise	Is your daily work physically straining	Ordinal	Distinguishes between 5 levels
Physical exercise	Exercise in the last 12 months	Ordinal	Distinguishes between 5 levels
SES	Education	Nominal	Illiterate Read/Write Primary Secondary Graduate

SES	Returning to Work	Nominal	On Sick Leave (SCT: 224459001) Unemployed (SCT: 73438004) Semi-Retired (SCT:224379008)
SES	Employment	Nominal	In paid employment (SCT: 406156006) Self-employed (SCT: 160906004) Unpaid work (SCT: 276061003) Unemployed (SCT: 73438004) Retired, life event (SCT: 105493001) Student (SCT: 65853000) Housemaid (SCT: 91534000)
Diet	Diet	Nominal	High fat diet (SCT: 226097005) Low fat diet (SCT: 16208003) Healthy diet (SCT: 226234005)
Blood Pressure/ Hypertension	Hypertension	Nominal	Yes No Unknown
Blood Pressure/ Hypertension	Peripheral pressure Central peripheral Pressure Peripheral pressure ratio Pulse wave velocity (PWV) LVEDP LVEDP Systolic blood pressure over 24h? LVEDP blood pressure over 24h? Systolic blood pressure during day Diastolic blood pressure during day Systolic blood pressure during night Diastolic blood pressure during night Radial mean arterial pressure	Numeric	Notes: PWV is a measurement of arterial stiffness; PWV increases with BP LVEDP: Left Ventricular end diastolic pressure
Stress	Stress	Nominal	Feeling stressed (finding) (SCT: 224974006) + Known absent (qualifier value) (SCT: 410516002) Feeling stressed (finding) (SCT: 224974006) + Known present (qualifier value) (SCT: 410515003) Feeling stressed (finding) (SCT: 224974006) + Unknown (SCT: 261665006)
Depression	Depression	Nominal	Complaining of feeling depressed (finding) (SCT: 272022009) + Known absent (qualifier value) (SCT: 410516002) Complaining of feeling depressed (finding) (SCT: 272022009) + Unknown (SCT: 261665006) Complaining of feeling depressed (finding) (SCT: 272022009) + Known

			present (qualifier value) (SCT: 410515003)
Sleep Disorder	Insomnia	Nominal	Present Absent Unknown
Diabetes	Diabetes Duration Insulin Level fP Glucose Wide range CRP	Numeric	
Atrial fibrillation	Size left atrium from heart echo examination Size septum from heart echo examination	Numeric	
Caratoid artery disease	Plaque from vascular examination, dexter Plaque from vascular examination, sinister	Boolean	
Cardiovascular Condition	Angina Major Cardiovascular Event Status Heart failure Atrial fibrillation Heart attack Stroke	Boolean	Each boolean feature records the presence or absence of the cardiovascular condition in the patient's medical record
Cardiovascular Condition	Systolic left ventricular function from echocardiography LVEDP left ventricular function from Doppler velocity Left ventricular function from Tissue Velocity Imaging Echocardiography of pulmonary veins during systole Echocardiography of pulmonary veins during diastole Enlarged heart (left ventricle divided by body surface)	Numeric	LVEDP: Left Ventricular end diastolic pressure
Dyslipidemia	Apolipoprotein A1 S-Triglycerides S-Cholesterol S-HDL Cholesterol S-LDL Cholesterol Ratio HDL/LDL	Numeric	
Thyroid Function	Parathyroid hormone Vitamin D	Numeric	Parathyroid hormone (regulates blood Ca conc) and Vitamin D (Regulates Ca and Ph)

Renal Function	S-Creatinine S-Potassium Cystatin Filtration rate Createnin clearing rate, GFR Albumin to Createnin Index (in Urine) Albumine (g/L)	Numeric	
Inflammation and infection	Interleukin-10 Interleukin-6	Numeric	
Genetic Factor	Angiotensin converting enzyme Genotype. Renin genotype	Nominal	Present Absent Unknown

3 Acute Stroke Treatment Models

This section focuses on the data schemas for the models designed for use after acute stroke events. The clinical use setting for models D4.6 and D4.7 is acute stroke treatment in a hospital setting. The first of these models D4.6 will predict short term stroke outcomes in terms of discharge NIHSS and/or modified Rankin Scale (mRS) 3 months post treatment. The NIHSS (National Institute of Health Stroke Scale) is a quantification of impairment caused by stroke, that assesses 15 items, including: level of consciousness, extraocular movements, visual fields, facial muscle function, extremity strength, sensory function, coordination (ataxia), language (aphasia), speech (dysarthria), and hemi-inattention (neglect). Similarly, the modified Rankin Scale provides a graded (7-level) evaluation of global disability post stroke. For more details on these scales (and related instruments) see deliverable D4.2. The clinical goal of this model is to help select the best therapy, in terms of surgery and patient specific medication, for acute stroke patients. This model will integrate medical imaging data and other forms of clinical data in its inputs. The second model D4.7 will predict a complex structured quality of life target profile for the patient (see Precise4Q deliverable D4.2 for a specification of these target outputs). The goal of this second model is to support the tailoring of therapeutic strategies to maximize the patient's long-term quality of life (i.e., to provide a perspective that extends beyond the shorter-term 3 months post treatment).

As before, for each model we identify a set of domain factors that are understood to be relevant to the clinical decision and then define a set of features that represent the factor in the model inputs. Note that in this section in some cases, most notably neuroimaging, a factor relates to an information modality as distinct from a biological domain factor. As was the case with model D4.5 the selection of the model's features is dependent on data availability. In this instance, the main data sources used in this feature selection process were the data used in Charité Universitätsmedizin Berlin (CUB) and the data from the Swedish Stroke Riksstroke register.

3.1 D4.6 Hybrid model predicting short term stroke outcome

Model D4.6 predicts short term stroke outcome. This model has a hybrid architecture (see Figure 2 below) that integrates two models:

1. a mechanistic simulation model of blood perfusion in the brain. This model takes vessel segmented 3D Neuroimaging as input and produces a 2D image of the Circle of Willis personalised to the individual patient;
2. and a phenomenological model (likely to be implemented using a deep learning architecture) which integrates the 3D Neuroimaging data, the 2D image of the Circle of Willis model generated by the mechanistic mode, and other clinical data to produce a prediction of the short-term stroke outcome for the patient.

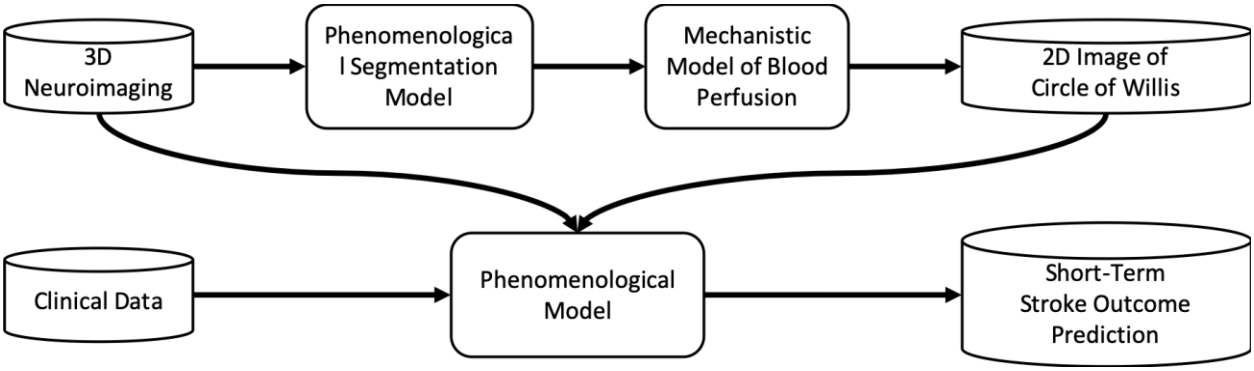

Figure 2 Schematic of the target architecture for model D4.6

It is frequently the case that in the acute setting the amount of data that is immediately available regarding a patient is relatively limited (as a minimum this might include *age*, *NIHSS*, and *imaging*). So we have taken an approach here where we have defined the feature set for this model as containing a relatively minimum set of features. What this means is that some factors that may be relevant to predicting the outcome for a patient, such as associated medical conditions (e.g. *diabetes*, *cardiac history*, *hypercholosterelomia* and so on) or medication (e.g. *antihypertensive agents*, *statins*, *platelet inhibitors* and so on) are not currently included. However, in the future we will explore whether extending the feature set to include these factors is worthwhile. Note that there is a strong correlation between disease status and medication (e.g. a patient with known hypertension and taking anit-hypertenstion medication) and so it may not be useful to include this information twice. Below we introduce the factors and features we currently include in the data schemas.

3.1.1 Neuroimaging

Neuroimaging is an important input modality for acute stroke treatment. Indeed, neuroimaging data is often the first type of data that a neurologist will consider when planning an acute stroke management strategy for a patient. Consequently, it makes sense to include neuroimaging as a primary form of input to the acute stroke models. Neuroimaging is performed on every patient

with stroke, primarily to distinguish patients suffering from ischemic stroke as distinct to brain bleed (haemorrhagic stroke). Both computed tomography (CT) and magnetic resonance imaging (MRI) are used. These images help to establish what proportion of stroke lesion volume is associated with stroke outcome (Thijs et al., 2000). Other parameters include presence of collaterals (Lu et al., 2019) or thrombus length (Rohan et al., 2014). Machine learning models, in particular deep learning models, can be used to integrate neuroimaging and clinical parameters. Another application of deep learning to neuroimaging is prediction of post-treatment infarction (Livne et al., 2018). Our acute models will use deep learning architectures to process and integrate neuro-imaging data with other types of data. Indeed, one of the advantages of deep learning architectures is the ability of these models to learn inter-modal representations that integrate information from different data-sources, be they image, text or structured data. Admittedly care must be taken when training these models to ensure that they attend to and learn from the different data sources. However, when done successfully these architectures have the potential to produce accurate predictions. Types of neuron imaging data that is currently used at Charité Universitätsmedizin Berlin (CUB) include: ***MRI SCAN, Angiography of head (CT ANGIOGRAPHY), Computed tomography Scan Brain (CT SCAN), Carotid Ultrasound.*** We are currently focusing on MRI imaging using the Time-of-Flight (TOF) MRA modality. TOF-MRA shows the structures of vessels which is useful for input to the mechanistic model of blood perfusion. Furthermore, it may be relevant to detecting the location and size of the lesion. We use a **voxel**-based representation for the scans: a voxel is a cube shaped region of brain tissue, with each voxel containing a million or so brain cells. A neurological image can be associated with a range of meta-data, for example timestamp of the imaging, the image modality, type (sequence), preprocessing steps, and radiological findings. It is expected that at least some of this metadata will be used as inputs to the models. In particular, the **radiological findings** will be included in the clinical data that will be input to the phenomenological model predicting the short-term stroke outcome. Examples of radiological findings include: *Occlusion*, *Lesion*, *Ischemia*, *Vessel malformation*, *Bleeding*, *Microangiopathy* and so on. Radiological findings may also be extended to include *territory* information, such as *frontal*, *left*, *occipital* and so on. Radiological findings data is often Boolean, or nominal. There may also be one or more **radiological scores** associated with a radiological finding. For example, a *lesion volume*. Typically, radiological scores are numeric.

3.1.2 Demographic Data and Obesity

Demographics data, such as **age** and **sex**, were identified as relevant factors in the prevention model. However, they are also relevant in predicting stroke treatment outcome. For example, age is the most well-known predictor of stroke outcome and stroke treatment success, as several studies have reported a direct association (Asadi et al., 2014; Khosla et al., 2010; Parsons et al., 2010; Weimar et al., 2002). Similarly, a number of studies have reported post-stroke outcomes

to be better for **obese** survivors of stroke, the so called 'obesity paradox' (Oesch et al., 2017). These factors are represented in the acute model inputs using the same features as were defined for the prevention model. See Table 2.1 for the operational definition of these features.

3.1.3 Time Interval Since Onset of Symptoms

The length of time between the onset of symptoms and treatment is a crucial factor in the outcome of a stroke treatment (Pulvers and Watson, 2017). For example, trails with moderate-dose intravenous thrombolytic treatments have shown that if administered within 3 hours after the onset of symptoms the treatment can have substantial benefits for patients, although there is evidence to support the use of this therapy beyond 3 hours (Ringleb P.A. et al., 2002). One complicating factor here is that the patient may have woken up with symptoms. To capture the time interval since onset we include two features in the current feature set, these are **Time Since Onset** a numeric feature recording the best estimate of the interval since the onset of symptoms recorded in minutes, and a boolean feature **Woke up with Symptoms**.

3.1.4 Stroke Type Diagnosis

The stroke diagnosis in terms of the type and the position and scale of injury is obviously a very important factor in predicting acute treatment outcome. We use a separate boolean flag to record each type of diagnosis. For example, we include a boolean flag for each of the following: ***Cerebral Haemorrhage. Cerebral Infraction, TIA, Acute Cerebrovascular Disease, Cerebral Haemorrhage with Ventricular Rupture,*** and so on.

3.1.5 Stroke Level Severity

There are a number of cognitive and motor function assessments that are known to be predictive of the severity of a stroke, and stroke outcome. For example, the 7-level modified Rankin Scale (**mRS**) for acute stroke is often used and provides a graded evaluation of global disability, mainly in terms of motor functions for a patient in daily living (Rankin, 1957). It is defined between 0 and 6, where an mRS of 0 indicates that no symptoms for disability are present, 5 denotes the most severe disabilities, and 6 records that the patient did not survive (van Swieten et al., 1988). The Fugl-Meyer Assessment (**FMA**) of neuromotor impairment following a stroke is another relevant assessment method. FMA is used for the assessment of motor function, balance, sensation and joint function in patients with post-stroke. This assessment helps record impairment of severity at motor and sensory functioning in both upper and lower extremities, balance while being seated and standing, and joint range of motion. However, at present we are using the National Institutes of Health Stroke Scale (**NIHSS**) as a quantification of impairment caused by stroke. It evaluates the initial neurologic outcome of stroke and the degree of recovery for patients with stroke (Schlegel et al., 2003, 2004). NIHSS is the most widely used and reliable state of the art

tool for the assessment of stroke level and thus a well-known stroke outcome predictor (Asadi et al., 2014; Parsons et al., 2010; Weimar et al., 2002). The scale assesses impairment for 15 items, including: level of consciousness, extraocular movements, visual fields, facial muscle function, extremity strength, sensory function, coordination (ataxia), language (aphasia), speech (dysarthria), and hemi-inattention (neglect) (Lyden et al., 2001, 1999). The scale ranges between 0 and 42 with higher values indicating a more severe stroke.

3.1.6 Treatment

The model also takes a representation of the type of treatment as an input. Currently we consider two types of treatment: **thrombolysis**, and **mechanical thrombectomy**. Thrombolysis is the most common treatment method and involves an agent being injected to dissolve the clot. Thrombolysis is mostly advised up to 4.5 hours post stroke (Lees et al., 2010). Mechanical thrombectomy has gained importance and clinical efficacy of this treatment has been reported in a number of studies, see: (Berkhemer et al., 2015; Campbell et al., 2015; Goyal et al., 2015; Jovin et al., 2015; Saver et al., 2015). In the current dataset there is an imbalance in the data with respect to the treatment type, with the majority of patients being treated via thrombolysis. Consequently, at present we represent treatment using a simple boolean feature that records true if thrombolysis was the treatment.

3.1.7 Table 2.1 Data Schema for Acute Stroke Treatment Models

Factor	Feature	Type	Notes, Example Values (SNOMED CT)
Neuroimaging	MRI Scan	Image	Modality: TOF-MRA, Representation: voxels
Neuroimaging	Radiological Finding	Nominal	Occlusion, Lesion, Ischemia, Vessel malformation, Bleeding, Microangiopathy
Neuroimaging	Radiological Score	Numeric	Lesion Volume
Demographic	Age (Patient's age)	Ordinal	Chronological age in 10-year (decile) ranges
Demographic	Sex (Patient's Biological sex)	Nominal	Female (SCT: 248152002) Male (SCT: 248153007)
Obesity	BMI	Numeric	BMI (SCT: 60621009)
Time Interval	Time Since Onset	Numeric	
Time Interval	Woke up with Symptoms	Boolean	

Diagnosis	Cerebral haemorrhage, Cerebral infarction, Transient ischemic attack, ...	Boolean	
Stroke Level Severity	NIHSS	Ordinal	0 No stroke symptoms 1-4 Minor stroke 5-15 Moderate stroke 16-20 Moderate to severe stroke 21-42 Severe stroke
Treatment	Thrombolysis	Boolean	

3.2 Personalised Acute Stroke Quality of Life Prediction Model (D4.7)

The target output for model D4.7 is an interrelated set of outcomes that attempts to capture a patient's overall Quality of Life (QoL). The goal of this model is to move beyond the estimates of the short-term disease outcome generated by D4.6 and instead generate a complex structured quality of life target profile for a patient. We have used the 2018 Riksstroke 3 Month Follow-up survey as the basis for our definition of this target output: specifically, model D4.7 will predict the patient's response to a number of the questions in this survey (see deliverable 4.2 for details). Consequently, given that the outputs for the model are defined relative to the 2018 Riksstroke 3 Month Follow-up survey it is natural that the input data for this model is also from the same data source. As a result, the primary data source for model D4.7 is the Swedish Stroke Riksstroke register. Given that model D4.6 and D4.7 are both acute models, natural there is overlap in relevant factors, however the definition of some of the features associated with the factors does vary between models D4.6 and D4.7 because of the differences in the data sources.

3.2.1 Neuroimaging

The Swedish Stroke Riksstroke register includes a number of features describing both CT and MRI scans. These include recording whether a scan was performed, whether the diagnosis showed a new cerebral infarction, whether angiography was performed and if so, which vessels were affected.

3.2.2 Demographic Data

As with model D4.6 demographic data is considered important in predicting outcomes. However, with respect to this factor the Riksstroke only records Gender as *man* or *woman*.

3.2.3 Time Interval Since Onset of Symptoms

As noted in D4.6 the length of time between the onset of symptoms and treatment is a crucial factor in the outcome of a stroke treatment. The Riksstroke registry records a number of features relating to time, including: whether the patient **Woke up with Symptoms** (yes, no, not known); the **Time of onset** (Hours.Minutes) of symptoms; and if Time of onset is unknown, it uses **Time interval from onset to arrival at hospital** which records the time interval since the most recent time the patient was known to be asymptomatic.

3.2.4 Stroke Type Diagnosis

The **stroke diagnosis** is recorded as one of four categories: *Cerebral haemorrhage*; *Cerebral infarction*; *Acute Cerebrovascular disease*, not specified as haemorrhage or infarction; and *TIA*. The data also records information on the **Site of cerebral haemorrhage**, and whether the Haemorrhage involved a **ventricular rupture**.

3.2.5 Stroke Level Severity

The Riksstroke register records two features related to stroke level of severity: the **Level of consciousness on arrival at hospital** and the **NIHSS** at admission (within 24 hours).

3.2.6 ADL/Accommodation Before Onset

The Riksstroke register provides data on a range of aspects of a subject's life prior to the onset of stroke. These include information relating to their accommodation and social supports, and their ability with respect to independent living as captured relative to activities of daily living. Admittedly, it is not likely that this data would be available in a standard acute setting, given that the output target of model D4.7 encompasses a relatively broad set of Quality of Life factors. Still, we explore whether a number of these features are helpful in predicting QoL. The details of these features are available in Table 3.1 below.

3.2.7 Associated Medical Condition (Risk Factors)

Diabetes, **atrial fibrillation**, **hypertension**, **smoking**, and **previous strokes** are all known risk factors of stroke and as such are likely to affect the QoL predictions for a patient post treatment. The Riksstroke register records information relating to these risk factors of stroke using nominal features (yes, no, not known). See Table 3.1 below for details of these features.

3.2.8 Swallowing Function/Speech Examination

Dysphagia is malfunctioning of oral cavity, pharynx, esophagus, or gastroesophageal junction causing difficulty in swallowing. Dysphagia is one of the many complications of stroke and is an independent marker of patient outcomes (Smithard, 2016). The Riksstroke register records a

number of features capturing whether a patient's **Swallowing function was tested**, whether the patient was **Swallowing function evaluated by a specialist** (either speech therapist or another dysphagia specialist), or **Speech evaluated by specialist**. Although these features do not record the actual outcomes of these assessment (these being reported in the patient's medical records) the fact that an assessment was carried out or not (and if so, why not), or that a specialist was present or not could be informative as to long-term outcomes.

3.2.9 Pharmaceutical Treatment

The Riksstroke register records data relating to a number of pharmaceutical treatments that may be administered at the onset of treatment. This includes **antihypertensive agents**, **statins**, **platelet inhibitors** and **oral anticoagulant**. The type of treatment applied to a patient often reflects the severity of the stroke and also a range of other contextual and other health factors of the patient and the stroke event. Consequently, the treatments applied during treatment can be predictive of outcomes. Interestingly, this data also records the '**main reasons for non-intervention with oral anticoagulants during treatment in the event of atrial fibrillation and heart infarction**' where some of these reasons include the patient having a *tendency to fall* or *dementia*. Capturing the fact that the patient has dementia is likely to be useful in predicting the patient's QoL.

3.2.10 Surgical Treatments

Two types of surgical treatment are currently reported in the Riksstroke data: **Hemicraniectomy**, and **Thrombectomy**. Hemicraniectomy is a surgical procedure where a large flap of the skull is removed and the dura is opened; this gives space for the swollen brain to bulge and reduces the intracranial pressure. Thrombectomy is a type of surgery to remove a blood clot from inside an artery or vein. For each of these types of surgery we can extract a feature from the Riksstroke data that records whether the surgery was performed, or if that status is unknown.

3.2.11 Data Schema for Acute Stroke Treatment Models

Note the specifications of many of the features in this table are based on version 18.a of the Riksstroke - Acute Phrase for Registration of Stroke.

Factor	Feature	Type	Notes, Example Values (SNOMED CT)
Neuroimaging	CT Scan During Treatment	Nominal	Yes, No, Unknown
Neuroimaging	CT Angiography Performed	Nominal	1a = yes, directly related to the initial CT scan 1b = yes, later during treatment 2 = no 3=examination within 28 days before onset of stroke 9=unknown

Neuroimaging	CT Angiography Performed of Affected Vessels	Nominal	1=carotid vessels 2=intracranial vessels 3=both carotid and intracranial vessels 9=unknown
Neuroimaging	MRI Scan During Treatment	Nominal	Yes, No, Unknown
Neuroimaging	MR Angiography Performed	Nominal	1a = yes, directly related to the initial CT scan 1b = yes, later during treatment 2 = no 3=examination within 28 days before onset of stroke 9=unknown
Neuroimaging	MR Angiography Performed of Affected Vessels	Nominal	1=carotid vessels 2=intracranial vessels 3=both carotid and intracranial vessels 9=unknown
Neuroimaging	Carotid ultrasound performed	Nominal	1=yes 2=no 3=examination within 28 days before onset of stroke 4=planned for after discharge 9=unknown
Demographic	Gender	Nominal	Man, Woman,
Time Interval	Time Since Onset	Numeric	
Time Interval	Woke up with Symptoms	Nominal	Yes, no, unknown
Time Interval	Time interval from onset to arrival at hospital	Nominal	1=within 3 hours 2a=within 4.5 hours 2b=within 6 hours 3=within 24 hours 4=after 24 hours 9=unknown
Diagnosis	Stroke Diagnosis	Nominal	I61=Cerebral haemorrhage I62=Cerebral infarction I64=Acute cerebrovascular disease (not specified as haemorrhage or infarction) G45.X=TIA
Diagnosis	Site of cerebral haemorrhage	Nominal	1=cerebrum, central/deep 2=cerebrum, lobar/superficial 3=cerebrum, unspecified if deep or superficial 4=brainstem 5=cerebellum 6=several different sites 7=other 9=not known
Diagnosis	Ventricular Rupture	Nominal	1=yes 2=no 9=not known

Stroke Level of Severity	NIHSS	Nominal	0 No stroke symptoms 1-4 Minor stroke 5-15 Moderate stroke 16-20 Moderate to severe stroke 21-42 Severe stroke
Stroke Level of Severity	Level of consciousness on arrival at hospital	Nominal	1 fully awake 2 drowsy but responding to stimulus 3 unconscious 9 not known
ADL/ Accommodation before Onset	Accommodation	Nominal	1=own accommodation without home help 2=own accommodation with home help 3=arrange accommodation 5=other 9=not known
ADL/ Accommodation before Onset	Living Alone	Nominal	1=patient lives entirely on his/her own 2=patient shared his/her household 9=not known
ADL/ Accommodation before Onset	Requires Assistance with ADL	Nominal	1=patient can cope without assistance 2=patient requires assistance 9=not known
ADL/ Accommodation before Onset	Mobility	Nominal	1=patient could move around without supervision both indoors and outdoors 2=patient was able to move around by himself/herself indoors but not outdoors 3=patient was assisted by another person when moving around/or was bedridden 9=not known
ADL/ Accommodation before Onset	Toilet Visits	Nominal	1=patient managed toilet visits without any help 2=patient was unable to get to bathroom or go to toilet without help 9=not known
ADL/ Accommodation before Onset	Clothes	Nominal	1=patient was able to get dressed without help 2=patient needed someone to fetch his/her clothes or needed help with dressing/undressed, or remained undressed 9=not known
Associated Medical Condition	A separate feature for each of the following: ● Previous Stroke ● Previous TIA ● Atrial fibrillation ● Diabetes ● Smoker ● Hypertension	Nominal	1=yes 2=no 9=not known
Swallowing Function/Speech Examination	Swallowing function tested	Nominal	1=yes 2=no/not known 3=not examined due to patient's reduced consciousness

Swallowing Function/Speech Examination	Swallowing function tested by Specialist	Nominal	1=yes 2=no need 3=no - need by no specialist available 9=not known or patient declined
Swallowing Function/Speech Examination	Speech evaluated by Specialist	Nominal	1=yes 2=no need 3=no - need but no specialist available 4=no - but order for after discharge 5=no 9=not known or patient declined
Pharmaceutical Treatment	A separate feature for each of the following: ● Antihypertensive agents ● Statins ● Platelet inhibitors ● Oral anticoagulant	Nominal	1=yes 2=no 3=no, planned intervention within 2 weeks after discharge 9=not known
Pharmaceutical Treatment	Reason for non-intervention with oral anticoagulants during treatment in event of atrial fibrillation and heart infarction	Nominal	1=planned after discharge 2=contraindications 3=interactions with other drugs/naturopathy 4=caution 5=tendency to fall 6=dementia 7=patient declined treatment 8=other reason 9=not known
Surgical Treatment	A separate feature for each of the following: ● Hemicraniectomy ● Thrombectomy	Nominal	1=yes 2=no 9=not known

4. Personalised rehabilitation model (D4.8) Data Schema

Stroke rehabilitation focuses on both cognitive and functional impairments that diminish a person's Quality of Life (QoL). During rehabilitation a patient's programme is regularly updated and revised: for example, the types of therapy offered, the intensity of the therapy, and the frequency and duration of therapy may be adjusted. Model D4.8 is designed to support this tailoring and updating of the rehabilitation by generating on a daily basis **a personalised rehabilitation schedule covering that day's activities**, as well as **forecasting the patient's cognitive and functional status on the final day of discharge**. The model will also predict the **Fragility (likelihood of a patient to retreat from life)** of a patient post rehabilitation. Note that it is likely that we will create different models to generate these different predictions: rehabilitation programme, forecasting day of discharge, and fragility. However, these different models will draw on the data schema defined below and can be integrated to provide the set of predictions specified for D4.8.

The Guttman Institute is the key Precise4Q partner dealing with stroke patient rehabilitation and so the features selected here are primarily based on the data available at Guttman. To date, the primary focus for discussions relating to rehabilitation programmes has been on cognitive rather than functional rehabilitation. This is partly driven by the fact that cognitive rehabilitation programmes are delivered at Guttman through computerized tasks (described below), meaning there is more data available in relation to cognitive rehabilitation. Consequently, although we discuss the functional assessment of patients and include these assessments as features in the data schema for the model, this is mainly to enable the model to predict the functional status on the final day of discharge; and so the features that are directly related to generating a rehabilitation programme are mainly connected to cognitive rehabilitation.

There are a number of factors described in previous models that are also included here, so we do not repeat the explanations of these factors for document brevity. For example, diagnosis (features: **diagnosis**), demographics (features: **age**, **sex, marital status**), obesity (feature: **BMI**). In addition, the time interval since the onset of symptoms is included, in this instance represented by counts of **Days Since Onset** and **Days in Rehab**.

4.1 Functional Assessment

There are a number of well-known instruments for the functional assessment of patients. There is, however, a large overlap between these instruments in terms of what they assess. For example, the Barthel and Functional Independence Scale assess a similar set of activities of daily living (ADLs). We briefly introduced ADLs previously in the section D4.7 Personalised Acute Stroke Quality of Life Prediction Model - ADL/Accommodation Before onset. Activities of Daily Living (ADLs) include the fundamental skills typically needed to manage basic physical needs, including: grooming/personal hygiene, dressing, toileting/continence, transferring/ambulating, and eating. At present we have chosen to use the Functional Independence Measure (FIM) instrument as our primary instrument for functional assessment (this choice was primarily driven by the data that is currently available). **FIM** assesses a patient across 18-items covering physical, psychological and social function, including: bowel and bladder control, transfers, location, feeding, grooming, bathing, and so on (Linacre et al., 1994). A patient is assessed on each item using a 7-point scale ranging from: 1 Total Assistance or Not Testable to 7 Complete Independence. FIM assessments are carried out when the patient is admitted and also at intervals throughout rehab. We include both the initial FIM assessments and the most recent FIM assessments in the data schema.

4.2 Cognitive Assessment

Post-stroke cognitive impairment is often reported in patients with stroke. For example, cognitive impairments can exhibit across a range of domains including memory, language, visuoconstruction, executive function, calculation, comprehension and judgment (Makin et al.,

2013). The assessment of a patient using the **NIHSS**[1] instrument includes assessment of many cognitive impairments and so we include the patient's NIHSS as part of the representation of cognitive assessment. The Guttman data also records a number of assessments focusing on specific cognitive functions, including: ***Level of Consciousness***, ***Orientation***, ***Attention***, ***Memory***, and ***Language.*** Stroke survivors are assessed on admission and also reassessed after each session of treatment to record the level of reduction in a particular deficit per specific function and also the overall deficit reduction. Consequently, these features are recorded both at admission and for the patient's most recent assessment.

4.3 Cognitive Rehabilitation Training

After accessing the stroke survivor's functional deficit in various neural functions, a cognitive rehabilitation (CR) program is designed and updated on a daily basis. The Guttman institute has developed the Guttman NeuroPersonalTrainer® platform (GNPT) for delivery of cognitive training through computerized tasks. The GNPT provides a library of tasks grouped by cognitive function that can be included in a patient's rehab programme. Note that the difficulty level of a task can be adjusted each time it is presented to a patient. Using the GNPT a neuropsychologist creates a daily CR treatment session by assigning a set of tasks to be completed that day. In addition, the difficulty of a task can be adjusted up or down each time a given task is presented to a patient. Furthermore, each time a patient completes a task, a task result score between 0 and 100 is calculated: the calculation of the task result score is dependent on the design of the task but the higher the score the better their performance on the task.

The combination of task selection, number of task repetitions, and task difficulty adjustment means that there are various treatment session configurations. The Guttman institute has developed the concept of ***Neurorehabilitation range (NRR)*** (García-Rudolph and Gibert, 2014) to guide treatment configuration. The NRR is the region within a CR task configuration space that produces maximum rehabilitation effects. The axes of a CR task configuration space are the number of executions of a task during a treatment session, and the performance (i.e. task result) in each execution of the task in the session. The motivation for the definition of the NRR is twofold: first, if the target performance level of a task is too low (i.e., too easy) then completing the task will only require the patient to use the undamaged areas of the brain, and so the impaired cognitive function will not be activated; conversely, if the target performance level is too difficult the impaired brain areas cannot respond to the difficult cognitive stimulus. The ideal is that a task configuration within a treatment session (in terms of repetitions and difficulty) is such that it is maximally beneficial to the patient in terms of reducing their deficit with respect to the cognitive function associated with the task. When a task is configured in this way it is considered to be in the NRR. The concept of the NRR allows for the dynamic adjustment of task

[1] The NIHSS is discussed in more detail in the acute models section where we discuss Stroke Severity Level.

difficulty within the GNPT. If a patient's task result is considered to be below the minimum threshold of the NRR then the difficulty level of the task is decreased at the next presentation; conversely, if a patient's task result is above the maximum threshold of the NRR then the difficulty level of the task is increased at the next presentation of the task. This difficulty adjustment is of course dependent on the specification of the NRR boundaries and this is one of the challenges that model D4.8 is designed to address.

Within the above context, we have developed a representation of the CR programme which includes both task specific features and treatment session features. For each task in the GNPT library we record the following:

- the total number of times the task has been assigned to the patient (**Count Task Assigned**),
- the number of days since task was last assigned to the patient (**Days Since Task Assigned**),
- the difficulty level for the most recent presentation of the task that resulted in a task result within the NRR (**Task NRR Difficulty**),
- the task result minimum threshold for the NRR for the most recent presentation of the task that resulted in a task result within the NRR (**Task NRR Min**)
- the task result maximum threshold for the NRR for the most recent presentation of the task that resulted in a task result within the NRR (**Task NRR Max**)

We also record the results of the last three sessions. Each session is represented by four vectors, each of these vectors is the same length as the number of tasks in the GNPT library. The first of these vectors records the number of times each task was included in the session (**Session Task Count**). The second vector records the highest difficulty level of the task in the session that resulted in the patient receiving a task result within the NRR (**Session Task Difficulty**). The third vector records the minimum task result threshold for NRR (**Session Task NRR Min**), and the fourth vector records the maximum task result threshold for NRR (**Session Task NRR Max**).

4.4 Fragility

One of the major challenges with stroke rehabilitation (and feeding into reintegration) is that after discharge from a rehabilitation programme patients can have a very low-level of compliance with their rehabilitation and retreat from life. One intervention that can help to avoid this is to provide patients with high-levels of support post-discharge, which can include home-help. At the Guttman institute the concept of **Fragility** (likelihood of a patient to retreat from life) is used to capture this dynamic: a patient that is deemed to be fragile at discharge is provided with extra support post discharge. However, these extra support services are expensive and must be targeted towards the most vulnerable patients. As a result, the Guttman institute has requested that a target relating to fragility be included within the rehabilitation model. A number of socioeconomical features are considered when assessing Fragility including: a patient's **level of Education**, whether they are **returning to work**, and **Employment**. Other features that are

considered include: the supports available in the **Accommodation** the patient is returning to, and whether they **Live Alone**; and patient's intrinsic resilience (in terms of **self-motivation, initiative,** and **compliance** during rehab).

4.5 Data Schema for Rehabilitation Model

Factor	Feature	Type	Notes, Example Values (SNOMED CT)
Diagnosis	Stroke Diagnosis	Nominal	I61=Cerebral haemorrhage I62=Cerebral infarction I64=Acute cerebrovascular disease (not specified as haemorrhage or infarction) G45.X=TIA
Demographic	Age (Patient's age)	Ordinal	Chronological age in 10-year (decile) ranges
Demographic	Sex (Patient's Biological sex)	Nominal	Female (SCT: 248152002) Male (SCT: 248153007)
Demographic	Marital Status	Nominal	Marital status: single, never married (SCT: 125725006) Cohabiting (SCT: 38070000) Divorced (SCT: 20295000) Widowed (SCT: 33553000) Separated (SCT: 13184001)
Obesity	BMI	Numeric	BMI (SCT: 60621009)
Time Interval	Days Since Onset	Numeric	
Time Interval	Days in Rehab	Numeric	
Cognitive Assessment	NIHSS (on admission and most recent assessment)	Nominal	0 No stroke symptoms 1-4 Minor stroke 5-15 Moderate stroke 16-20 Moderate to severe stroke 21-42 Severe stroke
Cognitive Assessment	Level of Consciousness (on admission and most recent assessment)	Nominal	fully conscious (SCT: 162701007) drowsy (SCT: 162704004) unconscious/comatose (SCT: 268913004)
Cognitive Assessment	Orientation (on admission and most recent assessment)	Nominal	Orientated (SCT: 247663003) Disorientated (SCT: 62476001)
Cognitive Assessment	Attention (on admission and most recent assessment)	Nominal	Able to direct attention (SCT: 288769005) Unable to direct attention (SCT: 288770006) Difficulty directing attention (SCt: 288773008)

Cognitive Assessment	Memory (on admission and most recent assessment)	Nominal	Temporary loss of memory (SCT: 162200009) Mild memory disturbance (SCT: 192071009) Memory function normal (SCT: 247602005) Amnesia (SCT: 48167000)
Cognitive Assessment	Language (on admission and most recent assessment)	Nominal	Able to use the elements of language (SCT: 288604009) Difficulty using the elements of language (SCT: 288608007) Unable to use the elements of language (SCT: 288605005)
Cognitive Rehabilitation	For each task in the GNPT library we record: ● Count Task Assigned ● Days Since Assigned ● Task NRR Difficulty ● Task NRR Min ● Task NRR Max	Numeric	
Cognitive Rehabilitation	For each of the last three sessions ● Session Task Count ● Session Task Difficulty ● Session Task NRR Min ● Session Task NRR Max	Numeric Vector	
Functional Assessment	FIM score for each of 18 functions at Admittance	Nominal	1 - Total Assistance or not Testable 2 - Maximal Assistance 3 - Moderate Assistance 4 - Minimal Assistance 5 - Supervision 6 - Modified Independence 7 - Complete Independence
Functional Assessment	FIM score for each of 18 functions at Most Recent Assessment	Nominal	1 - Total Assistance or not Testable 2 - Maximal Assistance 3 - Moderate Assistance 4 - Minimal Assistance 5 - Supervision 6 - Modified Independence 7 - Complete Independence

Fragility	Education	Nominal	Illiterate Read/Write Primary Secondary Graduate
Fragility	Returning to Work	Nominal	On Sick Leave (SCT: 224459001) Unemployed (SCT: 73438004) Semi-Retired (SCT:224379008)
Fragility	Employment	Nominal	In paid employment (SCT: 406156006) Self-employed (SCT: 160906004) Unpaid work (SCT: 276061003) Unemployed (SCT: 73438004) Retired, life event (SCT: 105493001) Student (SCT: 65853000) Housemaid (SCT: 91534000)
Fragility	Accommodation	Nominal	1=own accommodation without hom help 2=own accomodation with home help 3=arrange accommodation 5=other 9=not known
Fragility	Living Alone	Nominal	1=patient lives entirely on his/her own 2=patient shared his/her household 9=not known
Fragility	Resilience: ● Highly Motivated ● Strong Initiative ● High Compliance	Boolean	

5 Conclusions

This document sets out a set of data schemas for the prevention, acute, and rehabilitation models that are being developed in the PRECISE4Q project. These features included in these schemas have been extracted from stroke data available to us via to the consortium. It should be expected that as more data sources become available, and as data integration, harmonization, and model development progresses these data schemas will evolve.

References

1. Asadi, H., Dowling, R., Yan, B., Mitchell, P., 2014. Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy. PLoS ONE 9, e88225. https://doi.org/10.1371/journal.pone.0088225
2. Banerjee, A., Fowkes, F.G., Rothwell, P.M., 2010. Associations Between Peripheral Artery Disease and Ischemic Stroke: Implications for Primary and Secondary Prevention. Stroke 41, 2102–2107. https://doi.org/10.1161/STROKEAHA.110.582627
3. Béjot, Y., Bailly, H., Durier, J., Giroud, M., 2016. Epidemiology of stroke in Europe and trends for the 21st century. Presse Médicale 45, e391–e398. https://doi.org/10.1016/j.lpm.2016.10.003
4. Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., Delling, F.N., Djousse, L., Elkind, M.S.V., Ferguson, J.F., Fornage, M., Jordan, L.C., Khan, S.S., Kissela, B.M., Knutson, K.L., Kwan, T.W., Lackland, D.T., Lewis, T.T., Lichtman, J.H., Longenecker, C.T., Loop, M.S., Lutsey, P.L., Martin, S.S., Matsushita, K., Moran, A.E., Mussolino, M.E., O'Flaherty, M., Pandey, A., Perak, A.M., Rosamond, W.D., Roth, G.A., Sampson, U.K.A., Satou, G.M., Schroeder, E.B., Shah, S.H., Spartano, N.L., Stokes, A., Tirschwell, D.L., Tsao, C.W., Turakhia, M.P., VanWagner, L.B., Wilkins, J.T., Wong, S.S., Virani, S.S., On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee, 2019. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation 139. https://doi.org/10.1161/CIR.0000000000000659
5. Berkhemer, O.A., Fransen, P.S.S., Beumer, D., van den Berg, L.A., Lingsma, H.F., Yoo, A.J., Schonewille, W.J., Vos, J.A., Nederkoorn, P.J., Wermer, M.J.H., van Walderveen, M.A.A., Staals, J., Hofmeijer, J., van Oostayen, J.A., Lycklama à Nijeholt, G.J., Boiten, J., Brouwer, P.A., Emmer, B.J., de Bruijn, S.F., van Dijk, L.C., Kappelle, L.J., Lo, R.H., van Dijk, E.J., de Vries, J., de Kort, P.L.M., van Rooij, W.J.J., van den Berg, J.S.P., van Hasselt, B.A.A.M., Aerden, L.A.M., Dallinga, R.J., Visser, M.C., Bot, J.C.J., Vroomen, P.C., Eshghi, O., Schreuder, T.H.C.M.L., Heijboer, R.J.J., Keizer, K., Tielbeek, A.V., den Hertog, H.M., Gerrits, D.G., van den Berg-Vos, R.M., Karas, G.B., Steyerberg, E.W., Flach, H.Z., Marquering, H.A., Sprengers, M.E.S., Jenniskens, S.F.M., Beenen, L.F.M., van den Berg, R., Koudstaal, P.J., van Zwam, W.H., Roos, Y.B.W.E.M., van der Lugt, A., van Oostenbrugge, R.J., Majoie, C.B.L.M., Dippel, D.W.J., 2015. A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke. N. Engl. J. Med. 372, 11–20. https://doi.org/10.1056/NEJMoa1411587
6. Boehme, A.K., Esenwa, C., Elkind, M.S.V., 2017. Stroke Risk Factors, Genetics, and Prevention. Circ. Res. 120, 472–495. https://doi.org/10.1161/CIRCRESAHA.116.308398
7. Campbell, B.C.V., Mitchell, P.J., Kleinig, T.J., Dewey, H.M., Churilov, L., Yassi, N., Yan, B., Dowling, R.J., Parsons, M.W., Oxley, T.J., Wu, T.Y., Brooks, M., Simpson, M.A., Miteff, F., Levi, C.R., Krause, M., Harrington, T.J., Faulder, K.C., Steinfort, B.S., Priglinger, M., Ang, T., Scroop, R., Barber, P.A., McGuinness, B., Wijeratne, T., Phan, T.G., Chong, W., Chandra, R.V., Bladin, C.F., Badve, M., Rice, H., de Villiers, L., Ma, H., Desmond, P.M., Donnan, G.A., Davis, S.M., 2015. Endovascular Therapy for Ischemic Stroke with Perfusion-Imaging Selection. N. Engl. J. Med. 372, 1009–1018. https://doi.org/10.1056/NEJMoa1414792

8. Everson-Rose, S.A., Roetker, N.S., Lutsey, P.L., Kershaw, K.N., Longstreth, W.T., Sacco, R.L., Diez Roux, A.V., Alonso, A., 2014. Chronic Stress, Depressive Symptoms, Anger, Hostility, and Risk of Stroke and Transient Ischemic Attack in the Multi-Ethnic Study of Atherosclerosis. Stroke 45, 2318–2323. https://doi.org/10.1161/STROKEAHA.114.004815
9. García, A.M., Sedeño, L., Herrera Murcia, E., Couto, B., Ibáñez, A., 2017. A Lesion-Proof Brain? Multidimensional Sensorimotor, Cognitive, and Socio-Affective Preservation Despite Extensive Damage in a Stroke Patient. Front. Aging Neurosci. 8. https://doi.org/10.3389/fnagi.2016.00335
10. García-Rudolph, A., Gibert, K., 2014. A data mining approach to identify cognitive NeuroRehabilitation Range in Traumatic Brain Injury patients. Expert Syst. Appl. 41, 5238–5251. https://doi.org/10.1016/j.eswa.2014.03.001
11. Goyal, M., Demchuk, A.M., Menon, B.K., Eesa, M., Rempel, J.L., Thornton, J., Roy, D., Jovin, T.G., Willinsky, R.A., Sapkota, B.L., Dowlatshahi, D., Frei, D.F., Kamal, N.R., Montanera, W.J., Poppe, A.Y., Ryckborst, K.J., Silver, F.L., Shuaib, A., Tampieri, D., Williams, D., Bang, O.Y., Baxter, B.W., Burns, P.A., Choe, H., Heo, J.-H., Holmstedt, C.A., Jankowitz, B., Kelly, M., Linares, G., Mandzia, J.L., Shankar, J., Sohn, S.-I., Swartz, R.H., Barber, P.A., Coutts, S.B., Smith, E.E., Morrish, W.F., Weill, A., Subramaniam, S., Mitha, A.P., Wong, J.H., Lowerison, M.W., Sajobi, T.T., Hill, M.D., 2015. Randomized Assessment of Rapid Endovascular Treatment of Ischemic Stroke. N. Engl. J. Med. 372, 1019–1030. https://doi.org/10.1056/NEJMoa1414905
12. Haast, R.A., Gustafson, D.R., Kiliaan, A.J., 2012. Sex Differences in Stroke. J. Cereb. Blood Flow Metab. 32, 2100–2107. https://doi.org/10.1038/jcbfm.2012.141
13. Johnston, S.C., Mendis, S., Mathers, C.D., 2009. Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling. Lancet Neurol. 8, 345–354. https://doi.org/10.1016/S1474-4422(09)70023-7
14. Jovin, T.G., Chamorro, A., Cobo, E., de Miquel, M.A., Molina, C.A., Rovira, A., San Román, L., Serena, J., Abilleira, S., Ribó, M., Millán, M., Urra, X., Cardona, P., López-Cancio, E., Tomasello, A., Castaño, C., Blasco, J., Aja, L., Dorado, L., Quesada, H., Rubiera, M., Hernandez-Pérez, M., Goyal, M., Demchuk, A.M., von Kummer, R., Gallofré, M., Dávalos, A., 2015. Thrombectomy within 8 Hours after Symptom Onset in Ischemic Stroke. N. Engl. J. Med. 372, 2296–2306. https://doi.org/10.1056/NEJMoa1503780
15. Khosla, A., Cao, Y., Lin, C.C.-Y., Chiu, H.-K., Hu, J., Lee, H., 2010. An Integrated Machine Learning Approach to Stroke Prediction, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10. ACM, New York, NY, USA, pp. 183–192. https://doi.org/10.1145/1835804.1835830
16. Kim, W., Kim, E.J., 2018. Heart Failure as a Risk Factor for Stroke. J. Stroke 20, 33–45. https://doi.org/10.5853/jos.2017.02810
17. Kissela, B.M., Khoury, J., Kleindorfer, D., Woo, D., Schneider, A., Alwell, K., Miller, R., Ewing, I., Moomaw, C.J., Szaflarski, J.P., Gebel, J., Shukla, R., Broderick, J.P., 2005. Epidemiology of Ischemic Stroke in Patients With Diabetes: The Greater Cincinnati/Northern Kentucky Stroke Study. Diabetes Care 28, 355–359. https://doi.org/10.2337/diacare.28.2.355
18. Koo, D.L., Nam, H., Thomas, R.J., Yun, C.-H., 2018. Sleep Disturbances as a Risk Factor for Stroke. J. Stroke 20, 12–32. https://doi.org/10.5853/jos.2017.02887

19. Lees, K.R., Bluhmki, E., von Kummer, R., Brott, T.G., Toni, D., Grotta, J.C., Albers, G.W., Kaste, M., Marler, J.R., Hamilton, S.A., Tilley, B.C., Davis, S.M., Donnan, G.A., Hacke, W., 2010. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. The Lancet 375, 1695–1703. https://doi.org/10.1016/S0140-6736(10)60491-6
20. Linacre, J.M., Heinemann, A.W., Wright, B.D., Granger, C.V., Hamilton, B.B., 1994. The structure and stability of the functional independence measure. Arch. Phys. Med. Rehabil. 75, 127–132. https://doi.org/10.5555/uri:pii:0003999394903840
21. 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. https://doi.org/10.1161/STROKEAHA.117.019440
22. Lu, S., Zhang, X., Xu, X., Cao, Y., Zhao, L. bo, Liu, Q., Wu, F., Liu, S., Shi, H., 2019. Comparison of CT angiography collaterals for predicting target perfusion profile and clinical outcome in patients with acute ischemic stroke. Eur. Radiol. 29, 4922–4929. https://doi.org/10.1007/s00330-019-06027-9
23. Lyden, P., Lu, M., Jackson, C., Marler, J., Kothari, R., Brott, T., Zivin, J., 1999. Underlying structure of the National Institutes of Health Stroke Scale: results of a factor analysis. NINDS tPA Stroke Trial Investigators. Stroke 30, 2347–2354. https://doi.org/10.1161/01.str.30.11.2347
24. Lyden, P., Lu Mei, Levine Steven R., Brott Thomas G., Broderick Joseph, 2001. A Modified National Institutes of Health Stroke Scale for Use in Stroke Clinical Trials. Stroke 32, 1310–1317. https://doi.org/10.1161/01.STR.32.6.1310
25. Makin, S.D.J., Turpin, S., Dennis, M.S., Wardlaw, J.M., 2013. Cognitive impairment after lacunar stroke: systematic review and meta-analysis of incidence, prevalence and comparison with other stroke subtypes. J. Neurol. Neurosurg. Psychiatry 84, 893–900. https://doi.org/10.1136/jnnp-2012-303645
26. O'Donnell, M.J., Xavier, D., Liu, L., Zhang, H., Chin, S.L., Rao-Melacini, P., Rangarajan, S., Islam, S., Pais, P., McQueen, M.J., Mondo, C., Damasceno, A., Lopez-Jaramillo, P., Hankey, G.J., Dans, A.L., Yusoff, K., Truelsen, T., Diener, H.-C., Sacco, R.L., Ryglewicz, D., Czlonkowska, A., Weimar, C., Wang, X., Yusuf, S., 2010. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. The Lancet 376, 112–123. https://doi.org/10.1016/S0140-6736(10)60834-3
27. Oesch, L., Tatlisumak, T., Arnold, M., Sarikaya, H., 2017. Obesity paradox in stroke - Myth or reality? A systematic review. PloS One 12, e0171334. https://doi.org/10.1371/journal.pone.0171334
28. Pan, A., Sun, Q., Okereke, O.I., Rexrode, K.M., Hu, F.B., 2011. Depression and Risk of Stroke Morbidity and Mortality: A Meta-analysis and Systematic Review. JAMA 306, 1241. https://doi.org/10.1001/jama.2011.1282
29. Parsons, M.W., Christensen, S., McElduff, P., Levi, C.R., Butcher, K.S., De Silva, D.A., Ebinger, M., Barber, P.A., Bladin, C., Donnan, G.A., Davis, S.M., Echoplanar Imaging Thrombolytic Evaluation Trial (EPITHET) Investigators, 2010. Pretreatment diffusion- and perfusion-MR lesion volumes have a crucial influence on clinical response to stroke thrombolysis. J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab. 30, 1214–1225. https://doi.org/10.1038/jcbfm.2010.3

30. Paterson, K.E., Myint, P.K., Jennings, A., Bain, L.K.M., Lentjes, M.A.H., Khaw, K.-T., Welch, A.A., 2018. Mediterranean Diet Reduces Risk of Incident Stroke in a Population With Varying Cardiovascular Disease Risk Profiles. Stroke 49, 2415–2420. https://doi.org/10.1161/STROKEAHA.117.020258
31. Phua, C.S., Jayaram, L., Wijeratne, T., 2017. Relationship between Sleep Duration and Risk Factors for Stroke. Front. Neurol. 8, 392. https://doi.org/10.3389/fneur.2017.00392
32. Pulvers, J.N., Watson, J.D.G., 2017. If Time Is Brain Where Is the Improvement in Prehospital Time after Stroke? Front. Neurol. 8. https://doi.org/10.3389/fneur.2017.00617
33. Rankin, J., 1957. Cerebral Vascular Accidents in Patients over the Age of 60: II. Prognosis. Scott. Med. J. 2, 200–215. https://doi.org/10.1177/003693305700200504
34. Ringleb P.A., Schellinger P.D., Schranz C., Hacke W., 2002. Thrombolytic Therapy Within 3 to 6 Hours After Onset of Ischemic Stroke. Stroke 33, 1437–1441. https://doi.org/10.1161/01.STR.0000015555.21285.DB
35. Rohan, V., Baxa, J., Tupy, R., Cerna, L., Sevcik, P., Friesl, M., Polivka, J., Polivka, J., Ferda, J., 2014. Length of occlusion predicts recanalization and outcome after intravenous thrombolysis in middle cerebral artery stroke. Stroke 45, 2010–2017. https://doi.org/10.1161/STROKEAHA.114.005731
36. Rothwell, P.M., Algra, A., Amarenco, P., 2011. Medical treatment in acute and long-term secondary prevention after transient ischaemic attack and ischaemic stroke. The Lancet 377, 1681–1692. https://doi.org/10.1016/S0140-6736(11)60516-3
37. Sandsmark, D.K., Messé, S.R., Zhang, X., Roy, J., Nessel, L., Lee Hamm, L., He, J., Horwitz, E.J., Jaar, B.G., Kallem, R.R., Kusek, J.W., Mohler, E.R., Porter, A., Seliger, S.L., Sozio, S.M., Townsend, R.R., Feldman, H.I., Kasner, S.E., CRIC Study Investigators, 2015. Proteinuria, but Not eGFR, Predicts Stroke Risk in Chronic Kidney Disease: Chronic Renal Insufficiency Cohort Study. Stroke 46, 2075–2080. https://doi.org/10.1161/STROKEAHA.115.009861
38. Saver, J.L., Goyal, M., Bonafe, A., Diener, H.-C., Levy, E.I., Pereira, V.M., Albers, G.W., Cognard, C., Cohen, D.J., Hacke, W., Jansen, O., Jovin, T.G., Mattle, H.P., Nogueira, R.G., Siddiqui, A.H., Yavagal, D.R., Baxter, B.W., Devlin, T.G., Lopes, D.K., Reddy, V.K., du Mesnil de Rochemont, R., Singer, O.C., Jahan, R., 2015. Stent-Retriever Thrombectomy after Intravenous t-PA vs. t-PA Alone in Stroke. N. Engl. J. Med. 372, 2285–2295. https://doi.org/10.1056/NEJMoa1415061
39. Schlegel, D., Kolb, S.J., Luciano, J.M., Tovar, J.M., Cucchiara, B.L., Liebeskind, D.S., Kasner, S.E., 2003. Utility of the NIH Stroke Scale as a predictor of hospital disposition. Stroke 34, 134–137. https://doi.org/10.1161/01.str.0000048217.44714.02
40. Schlegel, D.J., Tanne, D., Demchuk, A.M., Levine, S.R., Kasner, S.E., 2004. Prediction of Hospital Disposition After Thrombolysis for Acute Ischemic Stroke Using the National Institutes of Health Stroke Scale. Arch. Neurol. 61, 1061–1064. https://doi.org/10.1001/archneur.61.7.1061
41. Smithard, D.G., 2016. Dysphagia Management and Stroke Units. Curr. Phys. Med. Rehabil. Rep. 4, 287–294. https://doi.org/10.1007/s40141-016-0137-2
42. Spackman, K.A., Campbell, K.E., Côté, R.A., 1997. SNOMED RT: a reference terminology for health care. Proc. AMIA Annu. Fall Symp. 640–644.
43. Thijs, V.N., Lansberg, M.G., Beaulieu, C., Marks, M.P., Moseley, M.E., Albers, G.W., 2000. Is early ischemic lesion volume on diffusion-weighted imaging an independent predictor of stroke

outcome? A multivariable analysis. Stroke 31, 2597–2602. https://doi.org/10.1161/01.str.31.11.2597

44. van Swieten, J.C., Koudstaal, P.J., Visser, M.C., Schouten, H.J., van Gijn, J., 1988. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 19, 604–607. https://doi.org/10.1161/01.str.19.5.604
45. Weimar, C., Roth, M.P., Zillessen, G., Glahn, J., Wimmer, M.L.J., Busse, O., Haberl, R.L., Diener, H.-C., on behalf of the German Stroke Data Bank Collaborators, o, 2002. Complications following Acute Ischemic Stroke. Eur. Neurol. 48, 133–140. https://doi.org/10.1159/000065512
46. Wilke, T., Groth, A., Pfannkuche, M., Harks, O., Fuchs, A., Maywald, U., Krabbe, B., 2015. Real life anticoagulation treatment of patients with atrial fibrillation in Germany: extent and causes of anticoagulant under-use. J. Thromb. Thrombolysis 40, 97–107. https://doi.org/10.1007/s11239-014-1136-8
47. Yan, H., Liu, B., Meng, G., Shang, B., Jie, Q., Wei, Y., Liu, X., 2017. The influence of individual socioeconomic status on the clinical outcomes in ischemic stroke patients with different neighborhood status in Shanghai, China. Int. J. Med. Sci. 14, 86. https://doi.org/10.7150/ijms.17241
48. Zhang, R., Zhong, C., Zhang, Y., Xie, X., Zhu, Z., Wang, A., Chen, C.-S., Peng, Y., Peng, H., Li, Q., Ju, Z., Geng, D., Chen, J., Liu, L., Wang, Y., Xu, T., He, J., 2019. Immediate Antihypertensive Treatment for Patients With Acute Ischemic Stroke With or Without History of Hypertension: A Secondary Analysis of the CATIS Randomized Clinical Trial. JAMA Netw. Open 2, e198103–e198103. https://doi.org/10.1001/jamanetworkopen.2019.8103