

PRECISE4Q



PREDICTIVE MODELLING IN STROKE

DELIVERABLE - RESUBMISSION

Project Acronym: **Precise4Q**

Grant Agreement number: **777107**

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

D2.1 – Overview of potential data sources and an operational plan to access available data sources

Revision: 2.0

Authors and Contributors	Catalina Martínez Costa (MUG); Jose Antonio Miñarro Giménez (UM); Nikola Lazovski (QMENTA); Paulo Rodrigues (QMENTA)		
Responsible Author	Catalina Martínez Costa	Email	catalina.martinez@medunigraz.at
	Beneficiary MUG	Phone	+4331638517880

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
1.1	21/10/19	Catalina MC	MUG	Initial draft
1.2	28/10/19	Nikola L	QMENTA	Risk management
1.3	25/10/19	Jose Antonio MG	UM	Risk management
1.4	29/10/19	Catalina MC	MUG	Final writing and review
2.0	30/10/19	Catalina MC	MUG	Final review

Date of delivery	Contractual:	31.10.2018	Actual:	31.10.2019
Status	final <input checked="" type="checkbox"/> /draft <input type="checkbox"/>			

Abstract (for dissemination)	This document describes the results of the data surveys which provide a detailed description of the datasets that will be used within the project. It also provides an overview on the ethics application process for each dataset and risk management estimations based on the level of data anonymization required by each partner and dataset.
Keywords	Data survey, ethics, risk management

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise.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



Table of Content

1	Introduction	5
1.1	Objectives	6
2	Description of the surveys	6
2.1	Data transfer survey	6
2.2	Detailed data survey	8
3	Results of the surveys	10
3.1	Data transfer survey results	10
3.2	Detailed data survey results	13
4	Ethical applications	25
5	Risk management and mitigation	25
6	Conclusions	26
	References	26

List of Figures

Figure 1	Data Transfer Survey: General Information	6
Figure 2	Data Transfer Survey: Technical aspects	7
Figure 3	Data Transfer Survey: Legal aspects	7
Figure 4	Detailed Data Survey: data elements grouped by stroke phase	8
Figure 5	Detailed Data Survey: detailed information about a data element	9
Figure 6	Detailed Data Survey: detailed information about a data element with coded values	9

List of Tables

Table 1	Technical aspects survey (Dataset 1 UOT)	10
Table 2	Legal aspects survey (Dataset 1 UOT)	10
Table 3	Technical aspects survey (Dataset 2 UOT)	10
Table 4	Legal aspects survey (Dataset 2 UOT)	11
Table 5	Technical aspects survey (CARDIPP LIU)	11
Table 6	Legal aspects survey (CARDIPP LIU)	11
Table 7	Technical aspects survey (RIKSSTROKE LIU)	11
Table 8	Legal aspects survey (RIKSSTROKE LIU)	12
Table 9	Technical aspects survey (Rehab dataset GUT)	12
Table 10	Legal aspects survey (Rehab dataset (GUT))	12
Table 11	Technical aspects (Reintegration GUT)	13
Table 12	Legal aspects survey (Reintegration GUT)	13
Table 13	Detailed data survey (Dataset 1 UOT)	15
Table 14	Detailed data survey (Dataset 2 UOT)	18
Table 15	Detailed data survey (CARDIPP LIU)	21
Table 16	Detailed data survey (Rehabilitation dataset GUT)	25
Table 17	Possible risks, their impact in the platform and mitigation and contingency measures...	26



Executive Summary

This document describes the results of the data surveys performed to get a detailed description of the datasets that will be used within the project. For each dataset we have carried out two surveys: (1) a data transfer survey to get an overview of the technical and legal aspects of data at each organisation and (2) a detailed data survey focusing on describing the data and its representation for harmonization purposes.

Chapter 2 describes the surveys and provides the answers of the surveys for each project dataset and partner.

Chapter 3 describes briefly the progress on the ethics application for each dataset.

Chapter 4 presents the Risk management estimations based on the level of data anonymization required by each partner and dataset.

Finally, some conclusions are provided.



1 Introduction

Within Precise4Q heterogeneous data from multidisciplinary sources will be integrated: genomics, microbiomics, biochemical; imaging including mechanistic biophysiological models of brain perfusion/function; social, lifestyle, gender; economic and worklife, requiring substantial efforts for information extraction, semantic labelling and standardisation.

Clinical data will cover all phases of stroke. The following table shows an overview of the datasets available for the project.

Source (Partner)	Patient number	Type	Data details	Language
AOK Nordost (AOK)	1.700.000	Insurance cohorts	Sociodemographic, diagnostic, treatment, procedural, cost data; Cost-benefit analyses	German
Estonian Genome Center (UOT)	52.000	Longitudinal	Sociodemographic, diagnostic, treatment, procedural, Omics Data, (Imaging data)	English Estonian
CloudRehab (GUT)	1000	Admissions	Sociodemographic, diagnostic, treatment data, clinical daily annotations	Catalan Spanish
Qvidlab (GUT)	250	Admissions	Sociodemographic, diagnostic, treatment data, long-term follow-up, community, integration	Catalan Spanish
GNPT (GUT)	290	Admissions	Sociodemographic, diagnostic, treatment data, Computer-based rehab program	Catalan Spanish English
PADRIAS/AquAS (GUT)	1000	Admissions	Sociodemographic, diagnostic, treatment data, Social-health care (visits)	Catalan
Riksstroke (LIU)	450.000	Registry	Sociodemographic, diagnostic, treatment, procedural data, follow-up data	Swedish English
UK Biobank (LIU)	500.000	Longitudinal	Sociodemographic, diagnostic, treatment, procedural, imaging data, Omics data	English
SCAPIS (LIU)	30.000	Longitudinal	Sociodemographic, diagnostic, treatment, procedural data, long-term data	Swedish English

In order to identify all potential data sources characteristics and consider all privacy and security requirements in this deliverable we provide a detailed inventory of some of the data sources, with descriptions of accessibility, size, content, formats, coding systems used, reliability, etc. This inventory will allow use case leaders and the machine learning team to:

- Select the segments of data and features of interest for each use case
- Specify the needs for data aggregation and typical queries to feed the predictive models
- Specify data quality requirements



1.1 Objectives

In order to characterize all data sources, we have created two data surveys to be answered by each partner for each dataset available for the project.

- Data transfer survey: provide an overview of the technical and legal aspects of the datasets at each organization.
- Detailed data survey: more detailed survey that focus on describing the data and how it is represented at each organization with harmonization purposes.

Here, we describe the results of both surveys, together with status of the individual ethical applications carried out by each organization. In addition, we provide the risk management estimations based on the level of data anonymization required by each partner and dataset.

2 Description of the surveys

2.1 Data transfer survey

The data transfer survey (https://data.qmenta.com/p4q/data_transfer.html) is divided into three main sections:

1. **General Information:** general information about the organization (Name; E-mail and Partner centre)

PRECISE4Q



PREDICTIVE MODELLING IN STROKE

:: Precise4Q Data Transfer Survey

Survey Description

This survey aims to get an overview of the technical and legal aspects of the datasets to be used within the Precise4Q project.

Instructions

Please, answer one survey for each dataset at your institution to be used within the Precise4Q project. You can either save/load the specification on/from our server if you were given tokens, or you can download the specification of the survey and sent to us.

Token:

General Information

Name:

E-mail:

Partner center:

Figure 1 Data Transfer Survey: General Information



2. Technical aspects: technical aspects of the dataset (storage format, degree of data structure, data schema, language, access information, data size)

Technical aspects

In what format do you have (store) the stroke data? Please select the options that apply.

Excel (or other similar spreadsheet-like formats)

JSON file

XML file

SQL Database (MySQL, PostgreSQL, Oracle etc.)

Other ...

What is the degree of structuring of your data?

Structured

Unstructured (free text)

Semi-structured

Images

Do you have a description of the data structure (i.e. data schema)?

Select option ...

What is the data language used?

Ex. English, French, German etc.

How can the data and the data schema specification (if possible) being accessed/downloaded?

What is approximately the size of the data?

Ex. 500MB, 1GB, 10GB, 1TB etc.

Figure 2 Data Transfer Survey: Technical aspects

3. Legal aspects: legal data aspects (access permission requirements, anonymization, sensitive data elements)

Legal aspects

Is a special permission or application form required to access/download the data?

Select option ...

Is data already anonymized?

Select option ...

Does any data location related jurisdiction apply?

Select option ...

Please explain details in the following field:

Can you enumerate any sensitive data elements? E.g. Full name, date of birth, etc.

Figure 3 Data Transfer Survey: Legal aspects



2.2 Detailed data survey

This survey (<https://data.qmenta.com/p4q/>) aimed to get detailed information about the stroke related data from each dataset and how it is represented at each site. Based on the Stroke summary proposed in deliverable 3.1, for each stroke phase (i.e. Prevention, Acute treatment, Rehabilitation, Follow-up) we propose a list of data elements or features with their respective value sets. E.g. for recording information about the patient *DIET* we propose the following predefined list of values: *HIGH FAT DIET*, *LOW FAT DIET*, *HEALTHY FAT DIET*. Both the list of data elements and their allowed value sets can be modified according to the specific dataset characteristics. In the following, a screenshot of the survey shows an excerpt of the list of data elements grouped by stroke phase and main headings such as “General Information”, “History”, etc. For each phase new data elements can be added.

Data Elements

Token:

<p>▼ PREVENTION PHASE DATA + Add new data element</p> <p>▼ GENERAL INFORMATION + Add new data element</p> <p>Data Element: SEX</p> <p>Data Element: ETHNICITY</p> <p>Data Element: AGE</p> <p>▼ LIFE STYLE + Add new data element</p> <p>Data Element: BMI</p> <p>Data Element: PHYSICAL EXERCISE</p> <p>Data Element: DIET</p> <p>Data Element: ALCOHOL DRINKING</p> <p>Data Element: SMOKING</p> <p>Data Element: DRUG ABUSE</p> <p>Data Element: MENTAL STRESS</p> <p>▼ HISTORY + Add new data element</p> <p>Data Element: CONDITION</p> <p>▼ CURRENT DISEASES / COMORBIDITIES + Add new data element</p>
--

Figure 4 Detailed Data Survey: data elements grouped by stroke phase

For each data element the proposed data type and list of values is shown, and the following detailed information can be provided (see Figure 5):

- If the data element is considered relevant for any of the project use cases
- If the data element is recorded in the corresponding dataset
- In case the element is recorded within the dataset, and is not represented as free text the name of the data element or its database path
- If the data element is multi-value
- Its corresponding data type (i.e. Free text, Integer, Decimal, Datetime, Date, Time, Boolean, Categorical data, Other)
- In case the Categorical data type is selected, then you can specify a coding system (e.g. ICD, LOINC, etc.)
- In addition to the coding system, you can specify the list of coded values allowed by your data element and mappings to the values we proposed (see Figure 6)
- Any additional comments



Name of the data element

SEX

Proposed data type and list of values for this type of element

CATEGORICAL**

- FEMALE
- MALE
- INDETERMINATE SEX

****Note:** This list might be incomplete or no representative, help us to improve it by specifying the allowed list of categories in your database.

Do you consider this data element important to be recorded for any of the project use cases?

YES

Do you record this data in your database?

YES

If your database is structured (no free text records), specify the Source/Database path or name of the data element?

Enter the source path/name

Is this data element MULTI-VALUE of the data type you have chosen?

NO

Data Type

Additional comments

Select option ...

FREE TEXT
 INTEGER
 DECIMAL
 DATETIME
 DATE
 TIME
 BOOLEAN
CATEGORICAL DATA
 OTHER

Figure 5 Detailed Data Survey: detailed information about a data element

Data Type

CATEGORICAL DATA

Please specify the coding system:

Ex: LOINC

Please, provide the allowed list of categories ([Add new category](#))

If you have a coding system specified, suggest mapping to concepts of that coding system.
 Also, if you can, suggest mapping to one of OUR proposed values.

Additional comments

Figure 6 Detailed Data Survey: detailed information about a data element with coded values



3 Results of the surveys

In the following we provide a summary of the results of each survey by partner.

3.1 Data transfer survey results

The following tables list the answers (or summary of answers) for the technical and legal aspects for each dataset at each organization. We provide the results for two datasets from the University of Tartu, Estonian Genome Center (UOT), for two datasets from the Linköping University (LIU) and another two datasets from the GUTTMANN institute.

AOK shared their schemas with us and due to strict local security requirements, their data will not be integrated into the data warehouse. However, CUB will be allowed to access it through a local AOK system.

University of Tartu (UOT) / Dataset 1

Technical aspects					
Data storage format	Structuring degree	Data description	Language	Data and schema access	Data size
Spreadsheet-like	Structured Unstructured Semi-structured	NO	English Estonian	Not known yet	10GB

Table 1 Technical aspects survey (Dataset 1 UOT)

Legal aspects					
Formal access permission	Access permission details	Anonymized	Data location jurisdiction	Details data location jurisdiction	Sensitive data elements
YES	Approval from Ethics Committee and application to the Estonian Genome Center.	YES	YES	Within EU cloud, authenticated secure servers, encrypted data transfer.	Personal ID codes and full names are removed.

Table 2 Legal aspects survey (Dataset 1 UOT)

University of Tartu (UOT) / Dataset 2

Technical aspects					
Data storage format	Structuring degree	Data description	Language	Data and schema access	Data size
Spreadsheet-like	Structured	NO	English Estonian	Not known yet	10GB

Table 3 Technical aspects survey (Dataset 2 UOT)



Legal aspects					
Formal access permission	Access permission details	Anonymized	Data location jurisdiction	Details data location jurisdiction	Sensitive data elements
YES	Approval from Ethics Committee and application to the Estonian Genome Center	YES	YES	Within EU cloud, authenticated secure servers, encrypted data transfer.	Personal ID codes and full names are removed.

Table 4 Legal aspects survey (Dataset 2 UOT)

Linköping university (LIU) / Dataset CARDIPP

Technical aspects					
Data storage format	Structuring degree	Data description	Language	Data and schema access	Data size
Spreadsheets-like	Structured	YES	Swedish	sent via email	3700kB

Table 5 Technical aspects survey (CARDIPP LIU)

Legal aspects					
Formal access permission	Access permission details	Anonymized	Data location jurisdiction	Details data location jurisdiction	Sensitive data elements
YES	Approval from Ethics Committee and application to LIU	YES	NO	-	NONE

Table 6 Legal aspects survey (CARDIPP LIU)

Linköping university (LIU) / Dataset RIKSSTROKE

Technical aspects					
Data storage format	Structuring degree	Data description	Language	Data and schema access	Data size
Spreadsheets-like	Structured	YES	Swedish	online	-

Table 7 Technical aspects survey (RIKSSTROKE LIU)



Legal aspects					
Formal access permission	Access permission details	Anonymized	Data location jurisdiction	Details data location jurisdiction	Sensitive data elements
YES	Approval from reg. ethics and application to registry	YES	YES	If personal data is to be processed at a location other than the principal research premises, a written personal data entry agreement must have been entered	NONE (excluded when applying)

Table 8 Legal aspects survey (RIKSSTROKE LIU)

GUTTMANN Institute (GUT) / Dataset Rehab phase

Technical aspects					
Data storage format	Structuring degree	Data description	Language	Data and schema access	Data size
Spreadsheets-like	Structured	YES	English	Not known yet	1M
Spreadsheets-like	Unstructured	YES	Spanish Catalan	Not known yet	10M

Table 9 Technical aspects survey (Rehab dataset GUT)

Legal aspects					
Formal access permission	Access permission details	Anonymized	Data location jurisdiction	Details data location jurisdiction	Sensitive data elements
By signed contract Regulation (EU) 2016/679	Approval from Ethics Committee and signed contract by each involved partner	YES	YES	Within EU cloud, authenticated secure servers, encrypted data transfer.	All personal elements have been removed, e.g.: Personal ID codes, full names, date of injury, date of birth

Table 10 Legal aspects survey (Rehab dataset (GUT))



GUTTMANN Institute (GUT) / Dataset Reintegration phase

Technical aspects					
Data storage format	Structuring degree	Data description	Language	Data and schema access	Data size
Spreadsheets-like	Structured	YES	English	Not known yet	1M
Spreadsheets-like	Unstructured	YES	Spanish Catalan	Not known yet	10M

Table 11 Technical aspects (Reintegration GUT)

Legal aspects					
Formal access permission	Access permission details	Anonymized	Data location jurisdiction	Details data location jurisdiction	Sensitive data elements
By signed contract Regulation (EU) 2016/679	Approval from Ethics Committee and signed contract by each involved partner	YES	YES	Within EU cloud, authenticated secure servers, encrypted data transfer.	All personal elements have been removed, e.g.: Personal ID codes, full names, date of injury, date of birth

Table 12 Legal aspects survey (Reintegration GUT)

3.2 Detailed data survey results

In the following we show a summary of the results of the detailed data surveys. We only provide the elements for those answered as relevant and present. We provide the results for two datasets from the Estonian Genome Center (UOT) and for the CARDIPP dataset at Linköping. For the Riksstroke dataset, the forms used to record the data are available at the Riksstroke registry web (<http://www.riksstroke.org/forms/>). We also provide the results for the GUTTMANN rehabilitation dataset.

University of Tartu (UOT) Dataset 1 (xjGcwHmPCAH37Yi)

Data Element	Name db path	Data type	Coding System	Value sets
PREVENTION PHASE DATA				
GENERAL INFORMATION				
SEX		CATEGORICAL		1 --> MALE 2 --> FEMALE
AGE		INTEGER		
LIFE STYLE				
BMI		DECIMAL		



PHYSICAL EXERCISE		CATEGORICAL		
DIET		CATEGORICAL		
ALCOHOL DRINKING		CATEGORICAL		
SMOKING		CATEGORICAL		
HISTORY				
CONDITION		CATEGORICAL	ICD-10	
CURRENT DISEASES / COMORBIDITIES				
CONDITION		CATEGORICAL	ICD-10	
FAMILY HISTORY				
CONDITION		CATEGORICAL	ICD-10	
ACUTE DISEASE PHASE DATA				
DIAGNOSIS				
DIAGNOSIS		CATEGORICAL	ICD-10	
CURRENT DISEASES / COMORBIDITIES				
CONDITION		CATEGORICAL	ICD-10	
DATE OF DIAGNOSIS		DATE		
PHARMACEUTICAL TREATMENT				
ANTIHYPERTENSIVE AGENTS		CATEGORICAL	ATC	
DATE OF PRESCRIPTION		DATE		
REASON FOR (NOT) PRESCRIPTION				
STATINS		CATEGORICAL	ATC	
DATE OF PRESCRIPTION		DATE		
REASON FOR (NOT) PRESCRIPTION				
PLATELET INHIBITORS		CATEGORICAL	ATC	
DATE OF PRESCRIPTION		DATE		
REASON FOR (NOT) PRESCRIPTION				
ORAL ANTICOAGULANT		DATE		
DATE OF PRESCRIPTION				



REASON FOR (NOT) PRESCRIPTION				
REHAB PHASE DATA				
SOCIAL BACKGROUND & EDUCATION				
LEVEL OF EDUCATION		CATEGORICAL		
MARITAL STATUS		CATEGORICAL		
OCCUPATION		CATEGORICAL		

Table 13 Detailed data survey (Dataset 1 UOT)

University of Tartu (UOT) Dataset 2 (Cz4OMYKbxeUefw8)

Data Element	Name db path	Data type	Coding System	Value sets
PREVENTION PHASE DATA				
GENERAL INFORMATION				
SEX	pat_sex	CATEGORICAL		N --> FEMALE M --> MALE
AGE		INTEGER		
LIFE STYLE				
BMI	KMI	DECIMAL		
PHYSICAL EXERCISE		TEXT		
ALCOHOL DRINKING		TEXT		
SMOKING	SUITS	BOOLEAN		
MENTAL STRESS		TEXT		
HISTORY				
CONDITION		CATEGORICAL	ICD10	
CURRENT DISEASES / COMORBIDITIES				
CONDITION		CATEGORICAL	ICD10	
DATE OF DIAGNOSIS		DATETIME		
FAMILY HISTORY				
CONDITION		TEXT		
HISTORY PROCEDURE UNDERTAKEN				
PROCEDURE		CATEGORICAL		
DATE OF PROCEDURE PERFORMED		DATETIME		
ACUTE DISEASE PHASE DATA				
DIAGNOSIS				



DIAGNOSIS		CATEGORICAL	ICD10	
SIDE OF THE LESION		TEXT		
DATE TIME ADMISSION HOSPITAL / STROKE UNIT		DATETIME		
DATE TIME DISCHARGE HOSPITAL / STROKE UNIT		DATETIME		
DATE TIME ARRIVAL HOSPITAL / STROKE UNIT		DATETIME		
ARRIVED BY AMBULANCE		TEXT		
DATE DECEASED		DATE		
WOKE UP WITH SYMPTOMS		TEXT		
CEREBRAL HAEMORRHAGE				
CEREBRAL HAEMORRHAGE CONDITION		CATEGORICAL		
HISTORY				
CONDITION		CATEGORICAL	ICD10	
CURRENT DISEASES / COMORBIDITIES				
CONDITION		CATEGORICAL	ICD10	
DATE OF DIAGNOSIS		DATETIME		
ADL / ACCOMODATION				
PATIENT LIVES ALONE		TEXT		
EXAMINATION OF BRAIN AND VESSELS				
CT SCAN				
COMPUTED TOMOGRAPHY SCAN BRAIN (CT SCAN)		CATEGORICAL		
DATE TIME THE CT SCAN WAS DONE		DATETIME		
RADIOLOGICAL DIAGNOSIS OF THE CT SCAN		TEXT		
MRI SCAN				
MAGNETIC RESONANCE OF BRAIN (MRI SCAN)		CATEGORICAL		
DATE TIME THE MRI SCAN WAS DONE		DATETIME		
RADIOLOGICAL DIAGNOSIS OF THE MRI SCAN		TEXT		
CT ANGIOGRAPHY				
COMPUTED TOMOGRAPHY ANGIOGRAPHY OF HEAD (CT ANGIOGRAPHY)		CATEGORICAL		



DATE TIME THE CT ANGIOGRAPHY WAS DONE		DATETIME		
RADIOLOGICAL DIAGNOSIS OF THE CT ANGIOGRAPHY		TEXT		
MR ANGIOGRAPHY				
MAGNETIC RESONANCE IMAGING ANGIOGRAPGY OF HEAD (MR ANGIOGRAPHY)		CATEGORICAL		
DATE TIME THE MR ANGIOGRAPHY WAS DONE		DATETIME		
RADIOLOGICAL DIAGNOSIS OF THE MR ANGIOGRAPHY		TEXT		
CAROTID ULTRASOUND				
CAROTID ULTRASOUND		CATEGORICAL		
DATE TIME THE CAROTID ULTRASOUND WAS PERFORMED		DATETIME		
RADIOLOGICAL DIAGNOSIS OF THE CAROTID ULTRASOUND		TEXT		
EXAMINATION OF HEART				
LONG TERM ECG		CATEGORICAL		
DATE TIME THE ECG WAS DONE		DATETIME		
ECG DIAGNOSIS		TEXT		
SWALLOWING FUNCTION / SPEECH EXAMINATION				
SPEECH DIFFICULTIES		TEXT		
SWALLOWING DIFFICULTIES		TEXT		
EVALUATION OF SPEECH FUNCTION		CATEGORICAL		
EVALUATION OF SWALLOWING FUNCTION		CATEGORICAL		
DATE TIME THE SWALLOWING EVALUATION WAS DONE		DATETIME		
PHARMACEUTICAL TREATMENT				
ANTIHYPERTENSIVE AGENTS		CATEGORICAL		
DATE OF PRESCRIPTION		DATETIME		
STATINS		CATEGORICAL		
DATE OF PRESCRIPTION		DATETIME		
PLATELET INHIBITORS		CATEGORICAL		
DATE OF PRESCRIPTION		DATETIME		
ORAL ANTICOAGULANT		CATEGORICAL		



DATE OF PRESCRIPTION		DATETIME		
TREATMENT / THROMBOLYSIS				
THROMBOLYSIS		TEXT		
SUBSTANCE ADMINISTERED		TEXT		
THROMBECTOMY				
THROMBOLYSIS		TEXT		
THROMBECTOMY SUBSTANCE		TEXT		
FOLLOW UP PHASE DATA				
TREATMENTS RECEIVED				
PROCEDURE		CATEGORICAL		
ACTIVITY LIMITATIONS AND PARTICIPATION RESTRICTION				
RETURN TO WORK		DATETIME		
ENVIRONMENTAL FACTORS				
PATIENT LIVES ALONE		TEXT		
REQUIRES ASSISTANCE		TEXT		

Table 14 Detailed data survey (Dataset 2 UOT)

Linköping university (LIU) Dataset CARDIPP

Data Element	Name db path	Data type	Coding System	Value sets
PREVENTION PHASE DATA				
Interleukin 10	IL10pgml	DECIMAL		
Interleukin 6	IL6pgml	DECIMAL		
Glipzid	glipizid_a	DECIMAL		
Metformin	metformin_a	DECIMAL		
globular filtrations-hastighet	MDRD_GFR_B	DECIMAL		
Arterial stiffness	kPWVcf_B	DECIMAL		
Body surface	vcKroppsyta_B	DECIMAL		
size septum	eSeptum_pr_a vnv_B	DECIMAL		
size left atrium	eLA_pr_avnv_B	DECIMAL		
Vitamine D	Vitamin_D	DECIMAL		
PTH	PTH	DECIMAL		
Albumine	ALBgL	DECIMAL		
IPHOS	IPHOSmmolL	DECIMAL		
Calcium level corrected for albumin	korrCa	DECIMAL		



Ionised calcium levels	CA_2mmolL	DECIMAL		
Angiotensin levels in plasma	AGT_SNP	DECIMAL		
Renin genotyp	Renin_genotyp	DECIMAL		
ATR1 genotype	ATR1_genotyp	DECIMAL		
genotype of the ACE complex	ACE_genotyp	DECIMAL		
Plaque	kPlaque_sum	DECIMAL		
Height of abdomen	kbukh	DECIMAL		
Pulmonary venous diastolic flow	eLungvenDiast	DECIMAL		
pulmonary venous systolic flow	eLungvenSyst	DECIMAL		
diastolic left ventricular function TVI	eDiastLVfunk_TV VI	DECIMAL		
diastolic left ventricular function	eDiastLVfunk_d oppler	DECIMAL		
Systolic left ventricular function	eSystLVfunk	DECIMAL		
Albumin/creatinin index	Ualb_Krea_index	DECIMAL		
lcte	lcte	DECIMAL		
Lipe	Lipe	DECIMAL		
APO B1	APO_B1	DECIMAL		
APO A1	APO_A1	DECIMAL		
CRP	wrCRP	DECIMAL		
GFR	CG_GFR	DECIMAL		
Insulin	S_Insulin	DECIMAL		
NtpBNP	P_NtpBNP	DECIMAL		
Cystatin	P_Cystatin_C	DECIMAL		
HBa1c	hba1c_IFCC	DECIMAL		
LDL cholesterol	ldl_kol_l	DECIMAL		
HDL cholesterol	hdl_kol_l	DECIMAL		
Cholesterol	S-Kolesterol	DECIMAL		
Triglycerides	tri_l	DECIMAL		
potassium	ka_l	DECIMAL		
creatinine	krea_l	DECIMAL		
glucose	glukos_l	DECIMAL		
thrombocytes	trombo_l	DECIMAL		
Leukocytes blood	leuko_l	DECIMAL		
Hemoglobin blood	hb_l	DECIMAL		
Overweight, obesity	synt12_p	CATEGORICAL		



Recurrent gastrointestinal disorders	symt11_p	CATEGORICAL		
Incontinence	symt10_p	CATEGORICAL		
Tinnitus	symt9_p	CATEGORICAL		
eczema, rash	symt8_p	CATEGORICAL		
sleeping problems	symt7_p	CATEGORICAL		
Tiredness	symt6_p	CATEGORICAL		
Anxiety	symt5_p	CATEGORICAL		
Headache or migrane	symt4_p	CATEGORICAL		
Pain in hands, elbows, legs or knees	symt3_p	CATEGORICAL		
Back pain, hip pain or sciatica	symt2_p	CATEGORICAL		
Pain in neck or shoulders	symt1_p	CATEGORICAL		
Blood pressure	bt1_sitt_syst_a bt1_sitt_diast_a bt2_sitt_syst_a bt2_sitt_diast_a bt3_sitt_syst_a bt3_sitt_diast_a bt_stående_syst_a bt_stående_diast_a	INTEGER		
Microalbumin	mikroalb_a	BOOLEAN		
GENERAL INFORMATION				
SEX	kön_a	CATEGORICAL		0 --> FEMALE 1 --> MALE
AGE		INTEGER		
LIFE STYLE				
Waist	midja_a	INTEGER		
BMI	BMI_VC_B, R_BMI_VC_B	DECIMAL		
PHYSICAL EXERCISE	anstr_1år_p, tid_varm_p, dagligt_arb_p			
ALCOHOL DRINKING	alko_1år_p, glas_typ_p, sex_glas_p, berusad_p	CATEGORICAL		
SMOKING	rökning_p	CATEGORICAL		1 --> NON SMOKER 2 --> NON SMOKER



				3 --> None
MENTAL STRESS	stressad_p	CATEGORICAL		1 --> NO 2 --> YES 3 --> YES 4 --> YES
HISTORY				
CONDITION	diab_debut_a, hypertoni_a, -, flimmer_a, synt12_p, stroke_a, hjärtsvikt_a (hjärtinfarkt_a, angina_a, flimmer_a, hjärtinfarkt_a)			
CURRENT DISEASES / COMORBIDITIES				
CONDITION		CATEGORICAL		
DATE OF DIAGNOSIS	infarkt_när_a, stroke_när_a, diab_debut_a	DATE		
FAMILY HISTORY				
CONDITION	heriditet_a	BOOLEAN		
HISTORY PROCEDURE UNDERTAKEN				
PROCEDURE	bypass_a			

Table 15 Detailed data survey (CARDIPP LIU)

GUTTMANN institute (GUT) Rehabilitation Dataset

SOURCE TERM	SOURCE TERM DATA VALUE	DATA TYPE	CODING SYSTEM
<i>GENERAL INFORMATION</i>			
SEX	FEMALE	CHAR	
	MALE		
	INDETERMINATE SEX		
	(SUBSET TO BE DEFINED)		
LATERALITY	LEFT HANDED vs RIGHT HANDED	CHAR	
AGE	INTEGER	INTEGER	
TIME SINCE ONSET OF STROKE	DATE TIME	INTEGER	
LENGTH OF STAY (REHAB)	DAYS	INTEGER	
BMI	DOUBLE	INTEGER	
<i>SOCIAL BACKGROUND & EDUCATION</i>			
LEVEL OF EDUCATION (PREVIOUS TO STROKE)	ILLITERATE	CATEGORICAL	
	READ/WRITE		
	PRIMARY		



	SECONDARY		
	GRADUATE		
MARITAL STATUS	SINGLE	CATEGORICAL	
	MARRIED		
	SEPARATED		
	DIVORCED		
	WIDOWED		
OCCUPATION	ACTIVE (PAID EMPLOYMENT)	CATEGORICAL	
	ACTIVE + PENSION		
	UNEMPLOYED		
	PENSIONIST		
	HOUSEMAID		
	STUDENT		
DIAGNOSIS			
DIAGNOSED CONDITION	(SUBSET TO BE DEFINED)	CHAR	
CONDITION SEVERITY	(SUBSET TO BE DEFINED FROM SUBCLASSES OF 272141005 Severities (qualifier value)	INTEGER	
SIDE OF THE LESION	(SUBSET TO BE DEFINED)	CHAR	
MEDICAL COMPLICATIONS	BLADDER DYSFUNCTION POST STROKE	CATEGORICAL	ICD9
	BOWEL DYSFUNCTION POST STROKE	CATEGORICAL	ICD9
	VENOUS THROMBOEMBOLISM POST STROKE	CATEGORICAL	ICD9
	SEIZURES POST STROKE	CATEGORICAL	ICD9
	OSTEOSPOROSIS POST STROKE	CATEGORICAL	ICD9
	CENTRAL PAIN STATES POST STROKE	CATEGORICAL	ICD9
	FATIGUE POST STROKE	CATEGORICAL	ICD9
	(SUBSET TO BE DEFINED)		
COGNITIVE REHABILITATION			
PROCEDURE	Cognitive Rehabilitation		
START DATE REHABILITATION	DATE TIME	DATE	
END DATE REHABILITATION	DATE TIME	DATE	
REHABILITATION ACTIVITIES	GNPT rehabilitations tasks (sessions, results)	CATEGORICAL	
PHYSICAL REHABILITATION			
PROCEDURE	Physical Rehabilitation		
START DATE REHABILITATION	DATE TIME	DATE	
END DATE REHABILITATION	DATE TIME	DATE	
REHABILITATION ACTIVITIES	OT (sessions), PT (sessions), complementary sessions	CATEGORICAL	



CURRENT LIST OF MEDICATION			
ANTIDEPRESSANTS	YES/NO	CHAR	
PAIN	YES/NO	CHAR	
ADL / ACCOMODATION			
IMPAIRMENTS OF BODY FUNCTIONS			
LEVEL OF CONSCIOUSNESS (NIHSS)	FULLY CONSCIOUS	INTEGER	
	UNCONSCIOUS		
	NOT KNOWN		
	DROWSY		
	(SUBSET TO BE DEFINED FROM On examination - level of consciousness)		
ORIENTATION (BATERIA)	ORIENTED	INTEGER	
	DISORIENTED	INTEGER	
ATTENTION (BATERIA)	ABLE	INTEGER	
	UNABLE	INTEGER	
	DIFFICULTY DIRECT ATTENTION	INTEGER	
MEMORY (BATERIA)	TEMPORARY LOS OF MEMORY	INTEGER	
	MILD MEMORY DISTURBANCE	INTEGER	
	MEMORY FUNCTION NORMAL	INTEGER	
	AMNSESIA		
LANGUAGE (BATERIA)	ABLE	INTEGER	
	UNABLE	INTEGER	
	DIFFICULTY USING THE ELEMENTS OF LANGUAGE	INTEGER	
MUSCLE POWER (NIHSS)	Insufficient power to move joint	INTEGER	
	No active muscle contraction	INTEGER	
	Movement against resistance	INTEGER	
	Movement against gravity	INTEGER	
	Visible muscle contraction only	INTEGER	
	Movement with gravity eliminated	INTEGER	
	Finding of grade of muscle power	INTEGER	
	Finding of Medical Research Council grade of muscle power	INTEGER	
	Muscle movement against resistance incomplete		
FEEL DEPRESSED (HIBS, PCRS)	YES	CHAR	ICD9
	NO	CHAR	ICD9
	UNKNOWN	CHAR	ICD9



PAIN (NURSE REGISTRES, NRS:0-10 AND LOCALIZATION OF PAIN)	YES	CHAR	ICD9
	NO	CHAR	ICD9
	UNKNOWN	CHAR	ICD9
ACTIVITY LIMITATIONS AND PARTICIPATION RESTRICTION (HEADING) FIM, BARTHEL			
COMMUNICATION	ABLE	INTEGER	
	UNABLE	INTEGER	
	DIFFICULTY COMMUNICATING	INTEGER	
WALKING	CAN MOVE WITHOUT HELP INDOORS AND OUTDOORS	INTEGER	
	CAN MOVE WITHOUT HELP ONLY INDOORS	INTEGER	
	IN BED OR ASSISTED BY SOMEBODY WHEN MOVING	INTEGER	
	CAN MOVE OUTSIDE	INTEGER	
	DEPENDENT FOR WALKING	INTEGER	
WASHING ONESELF	ABLE	INTEGER	
	UNABLE	INTEGER	
	DIFFICULTY TO WASH SELF	INTEGER	
TOILET VISITS	TOILET VISITS WITHOUT HELP	INTEGER	
	USED BEDPAN OR INCONTINENCE PADS	INTEGER	
	TOILET VISITS WITH HELP	INTEGER	
	NOT ABLE TO ADJUST CLOTHES	INTEGER	
DRESSING	ABLE TO GET DRESSED WITHOUT HELP	INTEGER	
	NEEDED HELP TO GET DRESSED	INTEGER	
EATING	ABLE	INTEGER	
	UNABLE	INTEGER	
	DIFFICULTY FEEDING SELF	INTEGER	
	NEEDS HELP WITH FEEDING	INTEGER	
FUNCTIONAL ORAL INTAKE SCALE	<p>TUBE DEPENDENT (levels 1-3)</p> <p>1 No oral intake</p> <p>2 Tube dependent with minimal/inconsistent oral intake</p> <p>3 Tube supplements with consistent oral intake</p> <p>TOTAL ORAL INTAKE (levels 4-7)</p> <p>4 Total oral intake of a single consistency</p> <p>5 Total oral intake of multiple consistencies requiring special</p>	INTEGER	



	preparation 6 Total oral intake with no special preparation, but must avoid specific foods or liquid items 7 Total oral intake with no restrictions		
<i>ENVIRONMENTAL FACTORS</i>			
ACCOMODATION AT DISCHARGE (ESIG-IG)	OWN ACCOMODATION WITHOUT HOME HELP	CATEGORICAL	
	OWN ACCOMODATION WITH HOME HELP	CATEGORICAL	
	ARRANGED ACCOMODATION	CATEGORICAL	
	HOSPITAL	CATEGORICAL	
PATIENT LIVES ALONE	LIVES ENTIRELY ON HIS/HER OWN	CATEGORICAL	
	SHARE WITH SPOUSE/PARTNER	CATEGORICAL	
REQUIRES ASSISTANCE	YES	CATEGORICAL	
	NO	CATEGORICAL	
	UNKNOWN	CATEGORICAL	

Table 16 Detailed data survey (Rehabilitation dataset GUT)

4 Ethical applications

The University of Tartu (UOT) has already initiated the research application for the Estonian Committee on Bioethics and Human Research and is ready to be signed by the project partners. After that an application to the Estonian Genome Center is required.

The University of Linköping has initiated the process to get the approval from the ethics committee for the CARDIPP dataset. For the Riksstroke dataset the ethics approval is ready to be signed by the corresponding partners. After the signature an application to the registry will be done.

GUTTMANN institute has shared with the partners a data processing agreement that has been already signed by some partners. After the signature they can access the data according to their roles (cf. deliverable 2.4).

As mentioned before, AOK due to strict local security requirements, will not allow to integrate their data into the project data warehouse. CUB, that is in the process of getting access, will be allowed to access AOK data through a proprietary AOK system.

5 Risk management and mitigation

Risk management and mitigation tasks are, mainly, mitigated by QMENTA Security Framework, which is described in more detail in deliverable 2.3 - Decision of build of the data warehouse, section 2.1 Security Framework; and the user account management using OAuth 2.0 [1], the industry-standard protocol for authorization, as described in deliverable 2.4.

When a user visits the QMENTA platform, communication channels are secured with HTTPS protocol [2] that encrypt all data that is exchanged with the platform. The access to a particular repository in the platform is allowed only for authorized users with their respective permissions (read, write, create, etc.) through username and password. Communications are monitored by QMENTA administrators to avoid attempts of unauthorized access and even to deactivate accounts. Furthermore, QMENTA retains audit logs to track the activity on the platform.



Based on the description above about security measures implemented in the platform where the Digital Stroke Patient Platform will be deployed, we have done an analysis of the risks. Table 12 shows a list of possible risks, their impact in the platform and the mitigation and contingency measures to implement.

Risk description	Probability	Impact	Mitigation and contingency measure
Unauthorized access to or action in the repository (cyberattacks)	Low	High	Administrators monitor the access to the data warehouse and control user interactions. In case of attempt of unauthorized access or unauthorized action, the user account will be deactivated until the person in charge of the account contacts the administrators.
Data loss or repository malfunction	Low	High	The Data warehouse is implemented into a cloud-based repository that encrypts the data and creates backups of the data repositories to protect them against catastrophic events. It provides for disaster recovery by allowing repository replication.
Communication interruption	Low	High	The platform is hosted on a cloud provider infrastructure and, therefore, it can be replicated into another data center to avoid communication problems.
Excessive computation resources	Medium	Low	Analysis modules can be executed in QMENTA platform and deployed using container technologies. The platform will allow to execute the modules outside the platform and only requiring access to the Data warehouse. In this case, the person responsible for the module development will be responsible for securing the datasets in their local system.

Table 17 Possible risks, their impact in the platform and mitigation and contingency measures

6 Conclusions

We have described the results of the data surveys performed to get a detailed description of the datasets that will be used within the project. For each dataset we have carried out two surveys: (1) a data transfer survey to get an overview of the technical and legal aspects of data at each organisation and (2) a detailed data survey focusing on describing the data and its representation for harmonization purposes. We have also addressed the current status of the application to the respective organization ethics boards in order to access the data. Finally, we have mentioned some possible risks related with unauthorized data access and malfunction of the repositories, together with their contingency measures.

References

1. OAuth 2.0. <https://oauth.net/2/> (accessed October 2019)
2. HTTPS protocol <https://support.google.com/webmasters/answer/6073543?hl=en> (accessed October 2019)