Funded by EU's Horizon 2020

D 9.1.
REVIEW OF EXISTING LONGITUDINAL POPULATION DATASETS

DISCLAIMER

ACKNOWLEDGEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.769661.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 769661

Supporting Active Ageing through Multimodal coaching

DELIVERABLE DOCUMENTATION SHEET

Deliverable:	*D 9.1*
WP №	*9*
Title:	*Review of Existing Longitudinal Population Datasets*
Editor(s):	*Senja Pollak, Saturnino Luz*
Contributor(s):	*Fasih Haider, Eleni Zarogianni, Pierre Albert, Sofia De La Fuente Garcia, Vera Veleva*
Type:	*Report/Documentation*
Version:	
Submission Due Date:	*31.7.2018*
Dissemination level (CO/PU):	*PU*
Copyright:	*©SAAM Project - All rights reserved*

☒ Approved by the WP Leader

☒ Approved by the Technical/Exploitation Manager[1]

☒ Approved by the Coordinator

☒ Approved by the PSC

[1] Choose Technical Manager for Deliverables in WP1-7,10 and Exploitation Manager in WP 8-10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

PUBLISHABLE SUMMARY:

Population health datasets are collections of data that contain health-related information (e.g., medical history, behavioural, genetic, phenotypic information) about large, possibly representative samples of their populations. The aim of SAAM is to develop methods for fine-grained sensor-driven longitudinal monitoring of ageing. The objective of this deliverable is to review a selection of population datasets that are related to ageing, including cognitive function, with the aim of assessing the availability of relevant data to be used in technological development of the project, and identifying opportunities for including data generated longitudinally by the SAAM technology into the existing datasets as well as for testing the SAAM technologies.

We analyse the characteristics of several datasets, including the information on demographics, biomarkers, clinical and cognitive testing, inclusion of sensor, speech and text data, data accessibility, possibility of data linkage, (secondary) data integration and collaboration including re-recruitment of participants.

The datasets included in this review are: UK Biobank, Generation Scotland, ADNI (Alzheimer's Disease Neuroimaging Initiative), ILSE (Interdisciplinary Longitudinal Study of Adult Development), SHARE (The Survey of Health, Ageing and Retirement in Europe), the DementiaBank Pitt Corpus and the Carolina Conversations Collection (CCC), which all have at least partial longitudinal monitoring. We also present the aggregation platforms UK Dementia platform, DementiaBank and several ongoing data gathering initiatives. While not exhaustive, this list covers the main characteristics of data relevant to SAAM.

The analysis shows that sensor-driven data is generally not included in the population datasets, with an exception of a large accelerometer monitoring dataset included in the UK Biobank study, that there are few examples of cognitive decline-related speech and text data, which can potentially be used in the development of methods for cognitive decline monitoring and that there are few datasets allowing for re-recruitment of participants and secondary data integration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

QUALITY CONTROL ASSESSMENT SHEET

Version	Date	Comment	Name of author/reviewer/contributor
V0.1	*20. 7. 2018*	*First Draft*	*Senja Pollak, Saturnino Luz, Eleni Zarogianni, Pierre Albert, Sofia De La Fuente Garcia, UEDIN* *Vera Veleva, BILSP*
	24. 7. 2018	*Review First Draft*	*Fasih Haider, UEDIN*
V0.2		*Contributions*	
	27.7.2018	*Second Draft*	*Senja Pollak, Saturnino Luz, Fasih Haider, UEDIN*
	28.7.2018	*1st Peer review*	*Zlatka Gospodinova*
	28.7.2018	*2nd Peer review*	*Nada Lavrač*
V0.3	*29.7.2018*	*Third draft*	*Senja Pollak, Saturnino Luz*
	30.7.2018	*WP Leader approval*	
	30.7.2018	*Coordinator approval*	
		EAB Review	*n.a.*
	31.7.2018	*PSC approval*	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

V1.0	*Submission to EC*

HISTORY OF CHANGES

For updating the Deliverable after submission to the EC if applicable

Version	Date	Change
V1.0		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Supporting Active Ageing through Multimodal coaching

PROJECT DOCUMENTATION SHEET

Project Acronym:	*SAAM*
Project Full Title:	*Supporting Active Ageing through Multimodal coaching*
Grant Agreement:	*GA № 769661*
Call identifier:	*H2020-SC1-2017-CNECT-1*
Topic:	*Personalised coaching for well-being and care of people as they age*
Action:	*Research and Innovation Action*
Project Duration:	*36 months (1 October 2017 – 30 September 2020)*
Project Officer:	*Dr. Reza RAZAVI*
Coordinator:	*Balkan Institute for Labour and Social Policy (BILSP)*
Consortium partners:	*Jožef Stefan Institute (JSI)* *University of Edinburgh (UEDIN)* *Paris-Lodron Universitat Salzburg (PLUS)* *Scale Focus AD (SCALE)* *Interactive Wear AG (IAW)* *Univerzitetni rehabilitacijski inštitut Republike Slovenije (SOČA)* *Nacionalna Katolicheska Federacia CARITAS Bulgaria (CARITAS)* *Bulgarian Red Cross (BRC)* *Eurag Osterreich (EURAG)*
website:	*saam2020.eu*
social media:	*#saam2020, #saamproject*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

ABBREVIATIONS

AD	Alzheimer's Disease
AMNART	American National Adult Reading Test
CDR	Clinical Dementia Rating scale
ADNI	Alzheimer's Disease Neuroimaging Initiative
CAPI	Computer-aided Personal Interviews
CCC	Carolina Conversations Collection
DPUK	Dementias Platform UK
EMCI	Early Mild Cognitive Impairment
GS	Generation Scotland
ILSE	Interdisciplinary Longitudinal Study of Adult Development
LMCI	Late Mild Cognitive Impairment
MCI	Mild Cognitive Impairment
SAAM	Supporting Active Ageing through Multimodal Coaching
SHARE	The Survey of Health, Ageing and Retirement in Europe

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

TABLE OF FIGURES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

I. INTRODUCTION AND MOTIVATION

Population health datasets are data collections containing health-related information (e.g., medical, genetic, family history, and lifestyle information) about large, possibly representative samples of their populations. The availability of large-scale longitudinal datasets and data linkage between different data collections opened new research directions in population health studies.

For this initial deliverable (to be updated in D9.2) we analyse a selection of datasets that are related to ageing and cognitive function: UK Biobank, Generation Scotland, ADNI (Alzheimer's Disease Neuroimaging Initiative), ILSE (Interdisciplinary Longitudinal Study of Adult Development), SHARE (The Survey of Health, Ageing and Retirement in Europe), the DementiaBank Pitt Corpus and the Carolina Conversations Collection (CCC), which all have at least partial longitudinal monitoring. We also present the UK Dementia platform, a data portal gathering information on more than 40 dementia-related cohorts, and the DementiaBank platform.

The selection covers datasets of various types, including biomedical datasets, either covering a specific disease (e.g., ADNI) or not bound to a specific disease, but intended to be used for a large number of prospective studies (e.g., UK Biobank, GS). We included also a multidisciplinary, large cross-European database (SHARE) and a collection of datasets including spontaneous speech samples (e.g., CCC).

The aim of SAAM is to develop methods for fine-grained sensor-driven longitudinal monitoring of ageing. The analysis of existing datasets gives initial information on the availability of population data that could be used in the SAAM technological developments (e.g., cognitive decline-related speech data) and opens perspectives on long-term integration or testing of SAAM technologies in the scope of longitudinal population datasets.

In Section II, we first provide a short description of each dataset, followed by an example of dataset aggregation platforms in Section III and examples of ongoing projects collecting relevant data (Section IV). The comparison of selected dataset is made in Section V. In Section VI, we briefly discuss the findings, comment on why and how the knowledge on the surveyed datasets could serve the SAAM project and, as the deliverable will be updated during the project, we sketch the plans for the future extensions of dataset descriptions.

II. DESCRIPTION OF SELECTED POPULATION DATASETS

For this initial deliverable (to be updated in D9.2) we selected the following population datasets: UK Biobank, Generation Scotland (GS), ADNI (Alzheimer's Disease Neuroimaging Initiative), ILSE (Interdisciplinary Longitudinal Study of Adult Development), SHARE (The Survey of Health, Ageing and Retirement in Europe), the DementiaBank Pitt Corpus and the Carolina Conversations Collection (CCC), which all have at least partial longitudinal monitoring.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

1. UK Biobank

UK Biobank (Sudlow et al. 2015, Allen et al. 2012) is a major health resource supported by the UK National Health Service (NHS). It is a prospective study following the health and well-being of 500,000 volunteer participants aged between 40-69 at the time of recruitment in 2006-2010.

The dataset contains extensive phenotypic and genotypic detail about its participants. Data results from physical measurements (e.g. blood pressure, grip strength, eye examination), questionnaires and interviews providing information on demographics, sociodemographics and family history, lifestyle (e.g., physical activity, smoking), psychosocial factors, environmental factors, health status (e.g., medications, disability), hearing threshold and cognitive function. Biological samples include blood, saliva and urine inventories for e.g., hormones and cholesterol testing etc. A subgroup of participants underwent imaging scans (brain, heart, abdomen, bones, carotid artery). A subset of around 100,000 UK Biobank participants have worn a 24-hour accelerometer (activity monitoring sensor) for a week (Doherty et al. 2017), providing UK Biobank with around 16 million hours' worth of data on physical activity.

The main goal is to study the relationship between genes, lifestyles and health, allowing researchers to study, for instance, why some people develop certain illnesses and others do not, and to measure the relative importance of "nature and nurture" in the development of illness. The data is linked to a wide range of electronic health records (cancer, death, hospital episodes, general practice). These health records are held by third parties, including NHS Digital (a division of the UK's National Health Service), the Information Services Division of NHS Scotland, and Public Health England (Follow-up of health). The dataset is available for health-related research in the public interest by researchers from the academic, charity, public, and commercial sectors, both in the UK and internationally, upon approval of application (which must contain a description of the research project in which the data will be used) and payment of a fee.

2. Generation Scotland

Generation Scotland (GS) was created to support medical research and identify the genetic basis of common complex diseases, including heart disease, diabetes, chronic pain and mental health (dementia, depression). The data collection is characterised by family-based recruitment, for analysing the association between genetic factors and a wide spectrum of illnesses and risk factors. GS is a partnership including 4 Scottish Universities (the University of Edinburgh, University of Glasgow, University of Aberdeen, University of Dundee), the NHS and people of Scotland.

The Generation Scotland concept has been evolving for several years. In total, over 30,000 people from across Scotland have contributed to the project, involving three complementary subprojects: Generation Scotland: the large genetic epidemiology resource Scottish Family Health Study (GS:SFHS), and smaller control samples, i.e. the Genetic Health in the 21st Century (GS:21CGH) and the Donor DNA Databank (GS:3D).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

- **GS:SFHS** (Smith et al. 2006, 2013) is the main cohort, designed as a family-based genetic epidemiology study of around 24,000 volunteers from around 7,000 families. The recruitment began in 2006 and ended in 2011. The participants answered questionnaires (gathering sociodemographic, lifestyle, personal and family health and medical history data), underwent clinical measurements (including anthropometric, cardiovascular respiratory, pressure testing) and cognition, mental health and personality assessment tests and questionnaires, and provided urine and blood or saliva samples. Biochemistry measures include Urea, Sodium, Creatinine, Glucose, Potassium, Total cholesterol, HDL cholesterol. For over 20,000 GS:SFHS participants a genome-wide chip genotyping (Illumina OmniExpress SNP GWAS, 700k) and exome chip (250K) data is available. The biological samples, together with broad phenotype and genotype data, form an important resource available for academic and commercial research. Broad consent was obtained for data linkage to medical records (routine NHS hospital, lab tests, prescribing and dental records, maternity and mortality data).

- Genetic Health in the 21st Century **GS:21CGH** is a resource of control DNA and genetic and phenotypic information, from nearly 2,000 consenting individuals, designed to help establish the genetic profile of a control population living in Scotland in relation to health and disease. The recruitment took place in 2007-2009. From each participant, two blood samples were collected (for DNA, plasma and peripheral blood leukocyte (PBL) testing), as well as a urine sample (not in all locations). Phenotype information was also collected, including basic physical observations, cognitive function measurements and answers to a clinical/lifestyle questionnaire.

- Donor DNA Databank **GS:3D** (Kerr et al. 2010) is a collection of DNA samples and plasma and some demographic information from the general healthy Scottish population, which is available to researchers investigating human diseases with a genetic component. GS:3D holds samples of around 5,000 volunteers, consented blood donors (aged 17-70 years). The purpose of GS:3D is to provide a long term, well characterised resource of human DNA control samples (for replication). The data was collected in 2008, DNA and plasma samples were unlinked and fully anonymised 28 days after collection. Only minimal phenotype information was collected.

GS is part of various consortia that focus on a variety of research areas such as ageing and mental health (Dementias Platform UK, European Prevention of Alzheimer's Dementia Consortium - EPAD, Stratifying Resilience and Depression Longitudinally - STRADL). No additional recruitment is foreseen, but GS allows re-contacting of participants, and it aspires to collect new data through online questionnaires and diaries, wearables and biometrics.

Also data linkage (for participants who have given consent) plays an important role in maximising the value of GS research data (raising of new research questions, replication of current findings), as it also allows continued follow-up of participants without requiring re-contact. The linkage uses the Community Health Index (CHI), which is Scotland's health care population register, the key to all patient data collected by the NHS in Scotland. Researchers can use the linked datasets to test research hypotheses on a stratified population and target recruitment to new studies. Future collaborations are planned for creating a combined cohort with UK Biobank and for inclusion of primary care and scanned

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

image data. The resources are available to academic and commercial researchers through a managed access process.

3. ADNI

ADNI (Alzheimer's Disease Neuroimaging Initiative) is a longitudinal, ongoing study, designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD).

It contains several overlapping cohorts/studies: ADNI 1, ADNI GO, ADNI 2 and ADNI 3:

- **ADNI1** (2004-2010): 400 subjects with MCI, 200 subjects with early AD and 200 elderly control subjects.
- **ADNI GO** (2009-2011): ADNI1 & 200 participants identified as having early mild cognitive impairment (EMCI). The objective of this phase was to examine biomarkers in an earlier stage of disease.
- **ADNI2** (since 2011): ADNI1 and ADNI GO cohorts & 150 elderly controls, 100 EMCI participants, 150 late mild cognitive impairment (LMCI) participants and 150 mild AD patients.
- **ADNI 3** (2016-2022): 697 from ADNI2 and 371 new (133 CN, 151 amnestic MCI, 87 AD)

The collected data contains (Demographics, Clinical Assessments, Cognitive Assessments), genetics, MR and PET imaging and imaging results, chemical and cognitive testing and chemical biomarkers.

Access is provided through LONI Image & Data Archive (IDA). After adherence to the ADNI Data Use Agreement and the publication policies, one can ask for ADNI imaging, clinical, genomic, and biomarker data for the purposes of scientific investigation, teaching or planning clinical research studies.

There are also related complementary dementia studies (also accessible through IDA): AIBL study (Australian Imaging Biomarkers and Lifestyle Study of Aging) and DoD-ADNI study (Effects of traumatic brain injury and post-traumatic stress disorder on Alzheimer's disease in Veterans using ADNI).

4. ILSE

Interdisciplinary Longitudinal Study of Adult Development - ILSE (Sattler et al. 2015) is an interdisciplinary longitudinal population-based study, created with the goal of studying individual, social and economic determinants of a healthy, self-determined and satisfied ageing.

The study investigates the aging process based on two German birth cohorts born between 1930-32 (C30) and 1950-52 (C50), respectively, which allows the opportunity to explore the potential impact of different childhood conditions (before and after World War II) on lifespan development. Moreover, the study sample includes participants from the former Western as well as Eastern (Leipzig region, LE) parts of Germany, allowing for the analysis of East/West Germany differences in developmental trajectories and outcomes. Given its first measurement wave at the beginning of the 1990s and its fourth

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

measurement wave, initiated in 2013, ILSE is now also able to contrast development in mid versus late adulthood.

ILSE uses the biographical approach, which emphasizes the importance of considering the full life course of individuals in order to understand their current motivations, concerns, and behaviours. The data contains detailed semi-standardized interviews, medical and psychogeriatric assessments, including the assessment of socio demographic characteristics (e.g., education, socioeconomic status), attitudes and values (e.g., toward the aging process, religion, politics), personality (e.g., Big Five assessment, control beliefs), indicators of subjective well-being (e.g., aging satisfaction, positive affect), social relationships (e.g., social support), as well as e.g., additional questions referring to physical involvement, media use, etc. ILSE's cognitively oriented and neuropsychological test battery was administered by professional psychologists.

As a result, ILSE contains over 8,000 hours of recorded biographic interviews from more than 1,000 participants over the course of 20 years, allowing investigations on various aspects of aging, such as cognitive decline, that often rely on the analysis of linguistic features. The dataset is of poor recording quality, audio segments are long, varying speaking styles and cross-talk, as well as emotional and dialectal speech. The methods for automatic transcription are being developed (Weiner et al. 2016) in order to enlarge the manually transcribed collection of 380 hours of ILSE interviews.

5. Pitt corpus

Pitt corpus is included in DementiaBank[2], a shared database of multimedia interactions for the study of communication in dementia that we describe in Section III. The Pitt corpus data was collected longitudinally, between 1983 and 1988, on a yearly basis; for a detailed description see Becker et al. (1994). Among the participants, there are around 200 who have Alzheimer's disease (AD) and around 100 are healthy adults. The dataset contains several subcorpora, generated according to neuropsychological tasks performed by the participants: the Cookie Theft description task, a Word Fluency task, a Story Recall task and a Sentence construction task, which are, depending on the task, performed either by healthy participant and dementia participants, or only by dementia participants. The Pitt corpus---and especially its Cookie Theft picture description task---has been used in many studies trying to predict or describe characteristics of AD based on speech or text data.[3] The participants are asked to describe what they see in a picture, and the material is in the form of dialogues between the interviewer and the participant. The transcriptions were all manually annotated with the CHAT guidelines, including the turns of participant/interviewer, repetitions, interruptions, errors, etc. (MacWhinney, 2000).

[2] https://dementia.talkbank.org/access/English/Pitt.html
[3] https://dementia.talkbank.org/publications/bib.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

6. Carolinas Conversations Collection (CCC)

Carolinas Conversations Collection (CCC, Pope and Davis 2011) is a US English spoken corpus consisting on conversations between adults older than 60 years and young interviewers, who were either students or community interviewers (medical, nursing or other health professionals), housed at the Medical University of South Carolina (MUSC). There are two cohorts in this dataset: one with 125 older speakers, from multiple ethnicities and any out of 12 chronic conditions, but no impairment; and another one with 125 speakers suffering from dementia, in a longitudinal set of approx. 400 conversations. Each participant gave two audio-recorded interviews that were recorded, talking about their health and their experiences with health care.

7. SHARE

The Survey of Health, Ageing and Retirement in Europe (SHARE)[4] is a multidisciplinary and cross-national database of micro data on health, socioeconomic status and social and family networks of more than 120,000 individuals aged 50 or older. SHARE covers 27 European countries and Israel.

The data collection started in 2004 and was gathered in more than 297,000 interviews (computer-assisted personal interviewing) complemented by measurements and written questionnaires. The collection has been organised in 7 waves[5]:

- **Wave 1** (2004) included 11 EU countries of various regions: Denmark, Sweden, Austria, France, Germany, Belgium, the Netherlands, Spain, Italy and Greece as well as Switzerland and Israel. The SHARE main questionnaire consisted of 20 modules on health, socio-economics and social networks. The data was collected by face-to-face, computer-aided personal interviews (CAPI), supplemented by a self-completion paper and pencil questionnaire.
- For **Wave 2** (2006-07), two 'new' EU member states (the Czech Republic and Poland) joint SHARE, as well as Ireland. The main addition in this wave was the 'End of Life' interview was conducted for family members of deceased respondents.
- **SHARELIFE** (2008-09) is the **Wave 3** of data collection for SHARE, which focuses on people's life histories (30,000 subjects). This dataset links individual micro data over the respondents' entire life with institutional macro data on the welfare state. With this variety SHARELIFE constitutes a large interdisciplinary dataset for research in the fields of sociology, economics, gerontology, and demography. The SHARELIFE life history data can be linked to the first two waves of SHARE assessing the present living conditions of older Europeans.

[4] http://www.share-project.org/

[5] http://www.share-project.org/data-documentation/waves-overview.html, https://en.wikipedia.org/wiki/Survey_of_Health,_Ageing_and_Retirement_in_Europe (Last accessed July 2018)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

- For **Wave 4** (2010-11) Estonia, Hungary, Luxemburg, Portugal and Slovenia joined the SHARE survey. A new social network module was added to the main questionnaire. In the German study, three additional projects including innovative biomarkers (e.g. dried bloodspots), the linkage with the German pension data as well as nonresponse experiments were implemented.
- **Wave 5** (2013) included 15 countries participated and included additional questions regarding childhood, material deprivation, social exclusion, and migration, as well as information on computer skills and the use of computers at the workplace.
- **Wave 6** (2015) was conducted in 17 countries. One of the most important innovations was the collection of objective health measures by means of "Dried Blood Spot Sampling" in 12 countries. Analyses were conducted for determining the blood levels associated with certain diseases (e. g., cardiovascular diseases, diabetes). These additional biomarkers are expected to be a useful instrument for comparing the objective health status with the subjective perception of the respondents. Moreover, they should help to explain correlations between health and social status and to demonstrate the course of a disease. Wave 6 furthermore captures longitudinal changes in the social networks.
- **Wave 7** collection (2017) took place in 28 countries - full coverage of the EU was achieved by including 8 new countries in SHARE: Finland, Lithuania, Latvia, Slovakia, Romania, Bulgaria, Malta and Cyprus. The Wave 7 questionnaire contains a SHARELIFE module for all respondents who did not participate in Wave 3 (first SHARELIFE wave), as well as a standard module for all respondents who already answered a SHARELIFE interview.

Upon registration, the data is available to the entire research community free of charge.

III. DATASET AGGREGATION PLATFORMS

In addition to datasets, there exist platforms that gather information about different datasets. As an example, we present the UK Dementia platform, a data portal gathering information on more than 40 dementia-related cohorts, and DementiaBank, an aggregation platform for dementia-related speech data.

1. Dementias Platform UK

Dementias Platform UK (DPUK)[6] links different cohorts of dementia with the aim to scale up dementia research within the UK. The contained datasets are listed in Table 1. It gathers data from around 2,000,000 study participants, with the aim of discovering the causes of dementia. An important tool is their cohort matrix comparison platform, by which over 40 datasets can be systematically compared. They also provide an online form to apply for access to these datasets and an Analysis Environment

[6] https://www.dementiasplatform.uk/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

(virtual desktop infrastructure) allowing to process and analyse the full range of physical, psychosocial and cognitive data (that is available in the cohort studies) on DPUK's servers without any physical transfer of data. The imaging platform serves a database for anonymised brain scans from different cohort studies. Researchers with granted access can manage, upload and share imaging data with the hub. Genetics platform is a collaborative platform to share and access research outputs from genetic research. Data Linkage service allows researchers to upload their own anonymised data into the Analysis Environment and use sophisticated, deterministic and probabilistic matching technologies to create de-identified, analysis-ready data.

Table 1: List of datasets linked in Dementia UK

A-D

- The Airwave Health Monitoring Study (**Airwave)**
- AMyloid imaging for Phenotyping LEwy body dementia (**AMPLE)**
- Brains for Dementia Research (**BDR)**
- Cambridge Centre for Ageing and Neuroscience (**Cam-CAN)**
- Cambridgeshire Parkinsons Incidence from GP to Neurologist (**CamPaIGN)**
- The Caerphilly Prospective Study (**CAPS)**
- Cognitive Function in Ageing Study II (**CFAS II)**
- The Cognitive Health in Ageing Register: Investigational, Observational, and Trial studies in dementia research (**CHARIOT)**
- CHARIOT: PRO Main Study (**CHARIOT: PRO)**
- Cognitive Health in Ageing Register: Investigational, Observational, and Trial studies in dementia research (CHARIOT): Prospective Readiness cOhort Study (PRO) (**CHARIOT: PRO Sub Study)**
- Project Cygnus (**Cygnus)**
- Dominantly Inherited Alzheimer Network (DIAN) Observational Study (**DIAN)**

D-J

- Emory Healthy Ageing Study (**EHAS)**
- Emory Healthy Brain Study (**EHBS)**
- The English Longitudinal Study of Ageing (**ELSA)**
- The European Prospective Investigation of Cancer - Norfolk (**EPIC Norfolk)**
- Environmental pollution-induced Neurological Effects (EPINEF) study (**EPINEF)**
- The GENetic Frontotemporal dementia Initiative (**GENFI)**
- Genetic and Environmental Risk in Alzheimer's Disease (GERAD) Consortium **(GERAD)**
- Generation Scotland: Scottish Family Health Study (**GS: SFHS)**
- The University of Hong Kong Neurocognitive Disorder Cohort (**HKUNCDC)**
- HealthWise Wales **(HWW)**
- The Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation-PD (**ICICLE-PD)**
- Lothian Birth Cohort 1936 (**LBC1936)**
- Identifying Predictors of dementia with Lewy bodies in People with Mild Cognitive Impairment (**LewyPro)**

K-N

- The Million Women Study (**Million Women)**
- Cognitive Function and Ageing Study (**MRC CFAS)**
- MRC National Survey of Health & Development (**MRC NSHD)**
- Northern Ireland Cohort for the Longitudinal study of Ageing (**NICOLA)**
- Neuroimaging of Inflammation in MemoRy and Other Disorders (the NIMROD Study) (**NIMROD)**
- Oxford Parkinson's Disease Centre Discovery Cohort (**OPDC Discovery)**
- Parkinson MRI Imaging Repository: Part 2 Database (**PaMIR)**
- Parkinsonism: Incidence and CogNItive heterogeneity in CambridgeShire (**PICNICS)**
- The PREVENT Research Programme (**PREVENT)**
- PRIME (etude PRospective sur l'Infarctus du MyocardE)
- Platform for Research Online To investigate gEnetics and CogniTion and ageing (**PROTECT)**
- Southall And Brent REvisited (**SABRE)**
- Samsung Medical Center Amyloid PET Cohort (**SMC Amyloid PET)**
- TRACK Huntington's disease (**TRACK HD)**
- UK Biobank
- Whitehall II

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

2. DementiaBank

DementiaBank is part of a larger TalkBank project (MacWhinney et al. 2011), gathered as part of the Alzheimer Research Program at the University of Pittsburgh. In addition to the already described **Pitt corpus** (see Section II), which was part of a larger longitudinal study, DementiaBank contains 8 other smaller corpora in English, German, Mandarin, Spanish and Taiwanese:

- **English Holland** (by Audrey Holland): videos of two individuals with Alzheimer's disease. The recordings contain language tasks from a Telerounds presentation (educational Grand Rounds dealing with different clinical speech and language problems of neurogenic origin).
- **English Kempler** (by Dan Kempler): dataset of conversation and Cookie Theft picture descriptions by six individuals with Alzheimer's disease.
- **English PPA DePaul** (by Roxanne DePaul): Primary Progressive Aphasia longitudinal data by 1 participant.
- **English PPA Hopkins** (by Argye Hillis): Primary Progressive Aphasia data for 36 participants (1/3 of participants were seen more than once). Diagnosis was based on cognitive and language testing and neurological examination and history, at Johns Hopkins Hospital.
- **German PPA** (by Fedor Jalvingh): Primary Progressive Aphasia data.
- **Mandarin Lu** (by Ching-ching Lu): dementia data of 52 participants.
- **Spanish PerLA** (by José Luis Mantero and Beatriz Gallardo-Pauls): Dementia data of 21 individuals with Alzheimer's disease.
- **Taiwanese Lu** (by Ching-ching Lu): Dementia data of 16 participants.

IV. DATASETS IN PROGRESS

There exist also several ongoing projects that have not yet released the collected datasets. For example, Toronto University's **Talk2Me** project[7] is a web-based and phone-based system (Talk2Me) for linguistic data acquisition, as means for longitudinal monitoring of changes in language ability. The data being gathered will be used to analyse the differences in longitudinal progression of linguistic markers between healthy older adults and the ones with dementia.

The **PREVENT** project aims to identify the earliest signs of dementia, which scientists believe may occur in the brain decades before symptoms appear. The participants are healthy volunteers in middle age (aged 40 – 59). The study is gathering data to identify biological and psychological factors that may increase their risk of developing dementia in later life, though a range of tests including blood tests, brain scans and cognitive assessments and follow how their brain health develops over time. The project

[7] https://www.cs.toronto.edu/talk2me/about/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

is for adding speech data collection in data collection (as part of the "Prevent triad" of behavioural data: speech, language and neuropsychological assessment).

The **EPAD** project[8] (The European Prevention of Alzheimer's Dementia project) is one of the largest dementia studies in the world. It is focussed on development of a novel, more flexible approach to clinical trials of drugs designed to prevent Alzheimer's dementia. Using an 'adaptive' trial design the results should be better and delivered faster and at lower cost. They are recruiting participants aged 50 or over, without dementia diagnosis. The participants will take part in clinical trials and regular assessments involving memory tests, brain scans and research samples (blood, saliva, spinal fluid and urine tests). Participants of the EPAD Longitudinal Cohort Study will be regularly monitored.

There are also several projects aiming at linguistic data collection, one example is the **Swedish Cookie-Theft Corpus,** which in being collected in the scope of the **Gothenburg MCI-study.** The Gothenburg mild cognitive impairment study (Wallin et al. 2016) conducts longitudinal in-depth phenotyping of patients with a wide range of cognitive impairment using neuropsychological, neuroimaging, and neurochemical tools. The study is clinically based and aims at identifying neurodegenerative, vascular and stress related disorders prior to the development of dementia. All patients in the study undergo baseline investigations, such as neurological examination, psychiatric evaluation, cognitive screening, magnetic resonance imaging of the brain and cerebrospinal fluid collection. At biannual follow-ups, most of these investigations are repeated. For speech data, a Swedish Cookie-Theft Corpus (Kokkinakis et al. 2018) is being collected, where the participants repeat the recording in 18-months time span, providing a longitudinal aspect. However, the dataset is not yet available.

V. ANALYSIS OF SELECTED POPULATION DATASETS

We analysed the datasets by several features, as presented in Tables 2, 3 and 4. We can see that collection of large medical population datasets started after year 2000 (UK Biobank, GS, ADNI), the same goes for the interdisciplinary SHARE dataset (starting in 2004). ILSE started in 1993, Pitt in 1983 and CCC in 2007.

The sizes of the selected datasets range from very large cohorts (UK Biobank with 500,000 participants), relatively big data collection (SHARE 120,000 individuals), followed by GS (30,000), ADNI and ILSE (around 1,000) and very small speech data samples from Pitt and CCC.

The age range at the time of recruitment for UK Biobank data is 40-69, for ADNI the age range for the study is 55-90, and for Pitt the age ranges between 46 and 89. CCC and SHARE describe the lower age limit (65 and 50, respectively), while ILSE defines two different cohort birth years 1930-32 and 1950-52. Generation Scotland (GS:SFHS) first recruited participants aged 35–65, but as the study is focused on family data, they were invited to engage at least one first-degree relative aged at least 18 years,

[8] http://ep-ad.org/about/project-objectives/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

resulting in final cohort age range of 18-98. The majority of datasets can be called longitudinal at least to some extent. While for some datasets the data collection process included several subject monitoring points, in other datasets the monitoring was performed only twice. Also, not all the participants in all the cohorts are monitored longitudinal (for some only data at a single point is available). The exception to longitudinal dataset is GS, but there the longitudinal aspect is assured by data linkage and re-recruitment possibilities.

Large population studies (UK Biobank, GS) are not limited to a specific disease but are used as data for prospective studies, for gathering information on a large number of diseases. Several selected datasets are on the other hand strongly related to Alzheimer's research (e.g., ADNI, Pitt as well as the Dementias Platform UK).

These datasets contain various types of data. In addition to more or less extensive sociodemographic data, participants provided biological samples (e.g., blood, urine, saliva) in UK Biobank, GS, ADNI, SHARE (for ILSE and Pitt, blood samples are reported but not part of the dataset). Genomics data is available in the UK Biobank, GS and ADNI. Imaging is characteristic for ADNI and UK Biobank, while medical measures are present in all datasets (unsurprisingly, the medical ones containing more detailed information), except for the two speech-based studies. Life style and activities data are also available (including information on diet, sleeping etc.) in a number of studies.

For neuro-cognitive testing, there are many standard examinations administered (MMSE, parts of Wechsler Memory Scale), Trail making, MoCa, ADAS-Cog, but also specifically developed tests (e.g., UK Biobank-designed cognitive tests administered through a computerised touchscreen interface).

CCC and Pitt contain spontaneous speech data as well as detailed transcripts, for ILSE part of the dataset is transcribed, while other dataset do not contain any speech or text (the only exception being possible availability of self-administered interview data, which is a very limited data source with relatively limited impact). The only real continuous sensor-driven monitoring dataset, incorporated into the above-described datasets, is a result of large accelerometer monitoring on 100,000 Biobank participants. There is also ongoing recruitment for heart monitoring.

Data availability is of crucial importance. From the perspective of SAAM researchers and other researchers developing methods for longitudinal sensor-driven monitoring data gathering and analysis, we identified as the most important the following characteristics of the surveyed datasets:

- **accessibility**: is a dataset accessible externally, outside the research group that created it?
  - motivation: researchers can use the data for training
- **openness for collaboration**: do dataset holders accept collaborations, re-recruit participants?
  - motivation: testing technologies in the scope of large population studies
- **extendibility**: is it possible to extend the existing dataset with new data/secondary data analysis?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

- motivation - new data: new data gathered by novel technologies (see collaboration above) is integrated into the existing datasets allowing other researchers to use it in large epidemiological studies showing e.g., correlation between sensor data and diseases
- motivation - secondary data: existing data processed by novel methods (e.g., derived features) can be re-used by large scientific community

- **linkage**: do dataset allow/provide linkage to other large datasets?
  - motivation: linked data from a dataset to large data repositories (e.g., information from visits of general practitioners, pension funds) can have large impact by revealing new, unknown links between diseases, genetics, societal factors and behaviour (e.g., patterns resulting from monitoring technologies).

UK Biobank data is available to all bona fide researchers for all types of health-related research that is in the public interest, regardless if it is performed by academia or commercial companies. An application review process is held and a fee charged prior to access being granted. Data reintegration is supported, as the researchers are asked to return their results to UK Biobank so that they are available for other researchers to use for health-related research that is in the public interest. UK Biobank is open for collaboration, inviting researchers to submit a proposal and enable to re-contact participants to collect new information via administration of a web-based questionnaire or a remote monitoring device.[9] UK Biobank has linked their data to national death and cancer registries and to national hospital data electronic record systems for all its participants since 2010. UK Biobank has also established linkages to primary care records from a large and ever-increasing proportion of its participants.

Generation Scotland has a described access procedure[10], inviting researchers to get in touch at an early stage of project planning. The application proposing a collaboration project is reviewed (including aspects, such as overlap with previous studies, ethics) and collaborators will be expected to meet additional costs. As consent for use the data out of the UK was not obtained at the beginning, the consents for SFHS are still being gathered. They are open to collaboration in terms of re-recruitment of participants. After the study, the researchers should provide the resulting data back to GS. In their promotional material, they explicitly mention activity monitoring data, diaries, as interesting future data extensions. GS data is linked to NHS records. The data on GS:SFHS participants was extracted by the NHS National Services Scotland electronic Data Research and Innovation Service (eDRIS), using the Community Health Index (CHI) number for linkage, then de-identified with a new ID.

ADNI has a transparent and light data access procedure[11], where the application including investigator's institutional affiliation and the proposed uses of the ADNI data should be submitted. The access is free of charge. All ADNI data is shared through the LONI Image and Data Archive (IDA), a secure research data repository. Interested scientists may obtain access to ADNI imaging, clinical, genomic, and

[9] http://www.ukbiobank.ac.uk/scientists-3/invitation-for-proposals-for-web-based-questionnaires-and-remote-monitoring-devices/
[10] https://www.ed.ac.uk/generation-scotland/using-resources/access-to-resources/access-process
[11] http://adni.loni.usc.edu/data-samples/access-data/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

biomarker data for the purposes of scientific investigation, teaching, or planning clinical research studies; ADNI data may not be used for commercial products. We did not find the information on other aspects of openness and data integration.

SHARE data is available and free of charge, but the registration should be made and the application must contain relevant details about the specific scientific project. The access to speech data CCC and Pitt is possible through a password-protected web site, provided that the researcher applies for access and agrees to comply with the provider's ground rules. For ILSE, the data access is not transparently described.

The TalkBank data through which the Pitt corpus is available is governed by the CC BY-NC-SA 3.0 license. Researchers can get password access by contacting the coordinator and becoming members. They should also briefly explain how the data will be used. In addition, by contact one can also get access to a specific version, which helps in the reproducibility of results. The integration of new and secondary data is possible, as researchers are asked whether they will be able to make the contribution to the database.[12] It is explicitly encouraged to contribute to TalkBank when creating new corpora (which might be relevant to SAAM, as new speech samples will be gathered).

For CCC the audio, video and transcript data are password-protected and stored on a secure server for use by approved researchers. Users whose research proposals meet ethical criteria of institutional review boards (IRBs) at home institutions and at MUSC may access the time-aligned transcripts, audio and video, for download and analysis using their own preferred tools, and may also analyse them online with Miner's suite of tools for acoustic signal, lexicon or syntax, downloading the results into spreadsheet. We did not find specific information with regard to collaboration and secondary data integration.

VI. CONCLUSION AND DISCUSSION

Population datasets are collections of genetic and phenotypic information on representative samples of their populations. They have been developed for monitoring population health-related issues in several countries (e.g., UK Biobank, ADNI in USA), with the goal of early identification of people at risk, discovering of disease-related risk markers across several genetic and non-genetic markers and other goals in diagnostics and research. We have analysed several datasets, including medical and large health related non-medical datasets (SHARE), as well as smaller datasets related to ageing, including cognitive decline, especially those that provide access to speech data (Pitt, CCC). We analysed the characteristics of the datasets, including the information on demographics, biomarkers, clinical and cognitive testing and availability of sensor, speech and text data.

[12] https://talkbank.org/share/rules.html

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

The analysis of existing datasets gives initial information on the availability of population data that could be used in technological development (e.g., cognitive decline-related speech data) and opens perspectives on long-term integration of data or testing of SAAM-technologies in the scope of longitudinal population datasets.

While there exist numerous sources of time series data reflecting people's daily activities, the described datasets do not include many of continuous monitoring data, which can be seen as a possible gap/missed opportunity in data gathering for population description. The few exceptions are a very large accelerometer data as part of the UK Biobank, an announced heart monitoring study (also UK Biobank), while GS explicitly mentions in their material that activity monitoring data, diaries, etc. would be future data extensions. The accelerometer data, as well as the described available speech collections from Pitt can be useful for SAAM partners for future technological developments.

We have assessed the datasets also from the perspective of accessibility, extendibility, openness for collaboration, extendibility and linkage. UK Biobank and Generation Scotland have transparent application procedure allowing for re-contacting of participants, which makes them good candidates for potential testing of developed technologies after the end of the project.

In further work, we will first extend the review of available datasets, by substituting the subjective dataset selection by a systematic query. We have searched the PubMed database by the following query:

- ((dataset[Title] OR data set[Title] OR datasets[Title] OR data sets[Title] OR biobank[Title] OR biobanks[Title] OR bio bank[Title] OR bio banks[Title] OR platform[Title] OR platforms[Title]) AND (aging[Title/Abstract] OR ageing[Title/Abstract] OR alzheimer*[Title/Abstract] OR dementia*[Title/Abstract] OR population[Title/Abstract] OR elderly[Title/Abstract] OR older[Title/Abstract] OR longitudinal[Title/Abstract] OR "cognitive decline" [Title/Abstract] OR "cognitive impairment" [Title/Abstract] OR "Alzheimer Disease"[Majr] OR "Cognitive Aging"[Majr] OR "Dementia"[Majr]))

which returned 2,279 results (on June 8 2018). We plan to systematically review the dataset described in the retrieved articles that will correspond to a set of predefined inclusion/exclusion criteria (longitudinal dataset related to ageing etc.). For the final collection, we will investigate to which extent large population datasets already incorporate the sensor-driven monitoring data.

For the updated Deliverable 9.2, we will study in more detail the possibilities for data integration into existing datasets. Especially from the perspective of identifying components of SAAM prototype technology, for which the re-recruitment of population dataset participants would allow designing large scale prototype testing of selected SAAM technologies beyond the scope of the project. We will also analyse in more detail, how the design of population datasets would best be done to facilitate the inclusion of sensor-driven data and specific challenges, including ethical aspects of data gathering and storage. We will consider making the proposed dataset analysis a living document, where the information about the relevant datasets will be updated by project partners, and reported on it in Deliverable 9.2.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Table 2: General datasets characteristics

Dataset	Timespan	Longi-tudinal	Participants P-patients, H-healthy	Age (recr.)	Diseases and mental health issues (S - specialized datasets)
UK BIOBANK	2006-2010 (recruitment)	Y	500,000 P, H	40-69	cancer, heart conditions, pain, eyesight/hearing conditions, disability, bipolar disorder, depression, anxiety, neurosis
GS Cohorts: GS:SFHS 21CGH GS3D	2006-2011 SFHS: 2006-2011 21CHG: 2007-2009 GS3D: 2008	N/Y through linkage	30,000 P, H SFHS: 24,000 21CHG: 2,000 GS3D: 24,000	18-98	heart disease, hypertension, stroke, diabetes, chronic pain; depression, anxiety
ADNI Cohorts: ADNI 1 ADNI GO ADNI 2 ADNI 3* *in progress	2004- 2004-2010 2009-2011 2011- 2016-	Y	> 1000 ADNI 1: 600 P, 200 H ADNI GO: adds 200 P ADNI 2: adds 400 P, 150 H ADNI 3: adds 238 P, 133 H	55-90	S: Alzheimer
ILSE Cohorts: C30: born 1930-32 C50: born 1950-52	1993-2016	Y	1002 P, H	~40,~60	psychiatric disorders, depression, medical history, physical assessment
PITT-DEMENTIA BANK	1983-1988	Y	~200 P, ~100 H	46-89	S: Alzheimer (and some comorbidities)
CCC	2007-2011	Y	300 (in data documentation) 46 from data set 30 P-AD, 16 NON-AD	>65	S: Alzheimer (Comorbidities: chronic diseases)
SHARE	2004-2017 7 waves	Y	120000 P, H	>50	Self-rated health, diseases, imputations, depression; end of life questionnaire (reasons and circumstances of death)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Table 3: Biomedical datasets characteristics

Dataset	Biomarker	Genomics	Imaging	Medical measures	Neurocognitive testing
UK BIO-BANK	blood, plasma, serum, buffy coat, urine, saliva, ...	~800,000 markers (SNPs and indels for 500,000 subjects) 25 genes screening: APOE, TREM2, … imputed: 500,000 WGS (planned for 50,000)	For subset: Brain MRI Heart MRI Bone DXA Abdomen MRI, carotid artery US scan	anthropometry, blood pressure, ECG, hand grip, spirometry, ECG cycle ergometry, eye examination, arterial stiffness, bone density (US), anamnesis	Biobank-designed cognitive tests (through computerised touchscreen interface): verbal and numeric reasoning, reaction time (to visual stimuli), short-term memory (numeric strings recall), visuospatial memory (pairs matching), prospective memory (the ability to remember and act on an instruction)
GS Cohorts: GS:SFHS 21CGH GS3D	SFHS: blood, serum, DNA, cryopreserved blood, urine, biochemical data 21CGH: blood, plasma, DNA, cells GS3D: DNA, plasma	SFHS: ~700,000 SNPs and 250,000 exome chip (20,000 subject) imputed (20,000) WES: ~1,000 WGS: ~20 APOE gene screening 21CGH: moderate genotyping	N	SFHS: anthropometry (incl. body fat), ECG, blood pressure, spirometry, anamnesis	Only for SFHS & 21CGH: WMS (selected) Digit Symbol test Verbal Fluency test Mill Hill Vocabulary Scale Eysenck Personality Questionnaire
ADNI Cohorts: ADNI 1 ADNI GO ADNI 2 ADNI 3*	blood, plasma, serum, CSF, cell immorta-lisation, RNA urine [ADNI 1]	ADNI GO/ADNI 2: ~700,000 SNPs WGS (~ 800 subj.) Gene screening: NCRAD, APOE, TOMM40 PolyT	Brain MRI: (fMRI/DTI), PET: PIB/FDG/Florbetapir	height physical exam, vital signs, respiration, pulse	MMSE, WMS (selected), Boston naming /MINT, Trail making, MoCA, ADAS-Cog AMNART, CDR, category fluency, clock drawing, Rey auditory/verbal learning test, financial capacity [ADNI 3], everyday Cognition-Participant Self Report [GO,2], ...
ILSE Cohorts: C30: born 1930-32 C50: born 1950-52	blood	APOE, COMT [wave 2,3]	Brain MRI [wave 2,3]	physical assessment anamnesis	MMSE, Hachinski, WMS (selected), Trail making, Boston naming Test, DSM-III-R (SKID), Bielefelder Autobiographical Memory Inventory, Nürnberger-Alters-Inventar, Leistungsprüfsystem,Aufmerk-samkeits-Belastungs-Test d2, Hamburg-Wechsler Intel. Test, depression screening (Zung scale)
PITT	blood (n.a.)	N	N	N	MMSE, Hachinki
CCC	N	N	N	N	N
SHARE 6 waves 1ongoing	dried blood spots [w6, 24,000],	N	N	anthropomet., grip str., respirat.,walk. speed, chair-stand	self-rated reading and writing skills, orientation, word list (immediate & delayed recall), verbal fluency & numeracy

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Table 4: Other dataset characteristics (sensor, speech, lifestyle, availability and linkage)

Dataset	Lifestyle, activities	Speech	Text	Continuous Monitoring Sensor Data	Linkage	Access	Recontact., secondary data integr.
UK BIOBANK	physical, smoking, drinking, diet, sleeping, ...	N	N	Accelerometer (~100000)	Y	Y application & fee	recontacting integration
GS Cohorts: GS:SFHS 21CGH GS3D	physical activities [SFHS]	N recordings for QA, destroyed after 1y	N (except from cog. test.)	N mentioned as future plan (wearables, online data, diaries)	Y	Y application & fee (outside UK subset of SFHS)	recontacting integration
ADNI	smoking, daily living activities FAQ	N	N	N	?	Y application free	N (?)
ILSE	physical, diet, leisure social support, media use, …	8,000 h biographic interview, poor rec. quality	transcripts of 380 hours of interview (erroneous)	N	N	linguistic data only - no clear procedure (by contact)	N (?)
PITT-DEMENTIA BANK	Blessed Activities Daily Living test for assessing changes	Y Cookie picture description	transcripts	N	N	Y free password-protected website	integration
CCC	N	Y dialogues (discussing health issues)	transcripts	N	N	Y application free password-protected website	N(?)
SHARE	leisure (voluntary work, clubs, relig. organ.), quality of life (CASP-12) ...	N	N (quest.)	N	Y SHARE-RV (Germany) REGLINK-SHAREDK (Denmark)	Y registration free	N(?)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

VII. REFERENCES

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., Sprosen, T. and Collins, R. 2012. UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology* 1(3): 123-126, doi.org/10.1016/j.hlpt.2012.07.003.

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J. and McGonigle, K. L. 1994. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology* 51(6): 585-594.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., van Hees, V. T., Trenell, M. I., Owen, C. G., Preece, S. J., Gillions, R., Sheard, S., Peakman, T., Brage, S. and Wareham, N. J. 2017. *PLOS ONE* 12(2): e0169649, doi: 10.1371/journal.pone.0169649.

Kerr, S. M., Liewald, D. C., Campbell, A., Taylor, K., Wild, S. H., Newby, D., … and Porteous, D. J. 2010. Generation Scotland: Donor DNA Databank; A control DNA resource. *BMC Medical Genetics* 11: 166, doi: https://doi.org/10.1186/1471-2350-11-166.

Kokkinakis, D., Lundholm Fors, K., Fraser, K. and Nordlund, A. 2018. A Swedish Cookie-Theft Corpus. In Calzolari et al. (eds.), *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, p. 1252-1258. Paris: ELRA.

MacWhinney B., Fromm, D., Forbes, M. and Holland, A. 2011. AphasiaBank: Methods for studying discourse. *Aphasiology* 25: 1286-1307.

MacWhinney B. 2000. *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W. and Beckett, L. 2005. The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America* 15(4): 869-877, doi:10.1016/j.nic.2005.09.008.

Pope, C. and Davis, B. H. 2011. Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory* 7(1): 143-161.

Sattler, C., Wahl, H.-W., Schröder, J., Kruse, A., Schönknecht, P., Kunzmann, U., … and Rahmlow, W. 2015. Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE). *Encyclopedia of Geropsychology,* p. 1213–1222.

Smith, B. H., Campbell, H., Blackwood, D., Connell, J., Connor, M., Deary I. J., Dominiczak A. F., Fitzpatrick B., Ford, I., Jackson, C., Haddow, G., Kerr, S., Lindsay, R., McGilchrist, M., Morton, R., Murray, G., Palmer, C.N., Pell, J.P., Ralston, S.H., St Clair D., Sullivan, F., Watt, G., Wolf, R., Wright, A., Porteous, D. and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Morris, A.D. 2006. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Medical Genetics* 7:74, doi: 10.1186/1471-2350-7-74.

Smith, B. H., Campbell, A., Linksted, P., McGilchrist, M., Wisely, L., Fitzpatrick, B., Ford, I., Hocking, L. J., Jackson, C., Kerr, S. M., Lindsay, R. S., Morton, R., Palmer, C. A. N., Deary, I. J., MacIntyre, D. J., Campbell, H., Dominiczak, A. F., Porteous, D. J. and Morris, A. D. 2013. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology* 42(3):689-700.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J. et al. 2015 UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* 12(3): e1001779. https://doi.org/10.1371/journal.pmed.1001779

Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., … Eckerström, C. 2016. The Gothenburg MCI study: Design and distribution of Alzheimers disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow & Metabolism* 36(1):114–131, doi: 10.1038/jcbfm.2015.147.

Weiner, M. W., Aisen, P. S., Jack, C.R., Jagust, W. J., Trojanowski, J. Q., Shaw L., Saykin, A. J., Morris J. C., Cairns, N., Beckett, L. A., Toga, A., Green, R., Walter, S., Soares, H., Snyder, P., Siemers, E., Potter, W,. Cole, P. E. et al. 2010. The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 6 (3): 202–211.e7. doi:10.1016/j.jalz.2010.03.007.

Weiner, J., Frankenberg, C., Telaar, D., Wendelstein, B., Schröderm J. and Schultz, T. 2016. Towards Automatic Transcription of ILSE — an Interdisciplinary Longitudinal Study of Adult Development and Aging. In Calzolari et al. (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016). Paris: ELRA.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661