D2.4 Identification of clinical data repositories

31/05/20

PERSIST (SC1-DTH-01-2019)

Project title	Patients-centered SurvivorShIp care plan after Cancer treatments based on Big Data and Artificial Intelligence technologies
Grant Agreement number	875406
Call and topic identifier	SC1-DTH-01-2019 - Big data and Artificial Intelligence for monitoring health status and quality of life after the cancer treatment
Funding schema	RIA
Coordinator	FUNDACION CENTRO TECNOLOXICO DE TELECOMUNICACIONS DE GALICIA (GRADIANT)
Website	www.projectpersist.com
Document keywords	clinical, non-clinical, data repositories
Document Abstract	Identifies data repositories containing clinical- and non-clinical datasets that might contribute to achieving the objectives of PERSIST and to contact these requiring restricted data Access.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

DOCUMENT

Authors	Simon LIN (SYMP), Roland ORTNER (SYMP)
Internal reviewers	Jean-Paul Calbimonte Pérez (HES-SO), Liliana Pires (RUBY)
Work package	2
Task	2.4
Nature	Report
Dissemination Level	Public

VERSION	DATE	CONTRIBUTOR	DESCRIPTION
0.1	2020-05-08	Roland ORTNER (SYMP)	First draft
0.2	2020-05-22	Jean-Paul Calbimonte (HES-SO)	Revision
0.**3**	2020-05-26	Liliana Pires (RUBY)	Revision
0.4	2020-05-27	Roland Ortner (SYMP)	Additions
0.5	2020-05-29	Roland Ortner (SYMP)	Revision
0.6	2020-05-29	Simon Lin (SYMP)	Revision
0.7	2020-05-31	Simon Lin (SYMP),	Final Version

DISCLAIMER

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

This document may contain material, which is the copyright of certain PERSIST consortium parties, and may not be reproduced or copied without permission. All PERSIST consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the PERSIST consortium as a whole, nor a certain party of the PERSIST consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk and does not accept any liability for loss or damage suffered by any person using this information.

ACKNOWLEDGEMENT

This document is a deliverable of PERSIST project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 875406

INDEX

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Executive Summary

This document lists all identified commercial or open datasets with clinical and non-clinical data that could be used during the project execution to feed the big data platform and artificial intelligence algorithms.

The identification of additional data repositories serve 2 purposes.

1. Complement existing data sources, which partners in the consortium bring into the project.
2. Risk mitigation to maximize data analysis outcomes and ensure usable results decision support and data extraction via subjective/objective sensors

Within the PERSIST project the goal is to develop impactful tools deviating patient trajectories of cancer survivors to the better. For this purpose electronic health records will be analysed using a data analysis pipeline which will be developed during the project, led by the project coordinator Grandiant and partners Symptoma, HESSO. Further a clinical decision support system will be developed by partners Emoda and DXC. Lastly a sensory network sperated in subjective and objective sensors will be created to capture patients health status.

For these 3 elements big data will be necessary and additionally to data brought in by partners, foremost clinical partners (SERVIZO GALEGO DE SAÚDE, Univerzitetni klinicni center Maribor, LATVIJAS UNIVERSITATE, CENTRE HOSPITALIER UNIVERSITAIRE DE LIEGE) following additional data repositories should ensure high yield results.

The following list of data sources was carefully curated evaluating 3 following points:

1. Data content and source
2. Access requirements
3. Estimation of impact on data analysis results on a scale 1-10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Introduction

Deliverable D2.4 identifies data-repositories containing **clinical datasets** that might contribute to achieving the objectives of PERSIST and to contact these requiring restricted data access. Some of the data repositories that were already identified are the NAMCS/NHAMCS - Ambulatory Health Care Data Homepage (CDC), the Cancer Imaging Archive, the CancerDR on drug resistance and the Genes-to-Systems Breast Cancer (G2SBC) Database, which contains mathematical models of carcinogenesis, tumour growth and response to treatments.

Moreover, this deliverable identifies **non-clinical datasets** that can be used to deliver digital biomarkers (e.g. night/day activity, facial, acoustics and text markers) of mood that can help researchers identify symptoms of mental illnesses (depression and anxiety syndrome). The non-medical datasets will also be used to deliver the conversational interface of the mobile application for patients.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Overview Identified data repositories

The following table provides an overview of the collected data repositories identified, the type of data (clinical or non-clinical) and information about access restrictions:

Name / Link	Type	Access
National Ambulatory Medical Care Survey (NAMCS)/ National Hospital Ambulatory Medical Care Survey (NHAMCS)	clinical	Open
Genes-to-Systems Breast Cancer (G2SBC) Database	clinical	Open
National Statistics Institute (INE)	clinical, non-clinical	Open
Medical Information Mart for Intensive Care III (MIMIC-III)	clinical	Free access
MTSamples	clinical	Open
Medical Records 10 yrs	clinical	Paid/Free
Registry of Specialized Health Activity (RAE-CMBD)	clinical	Open
The Depresjon Dataset	clinical	Open
Distress Analysis Interview Corpus (DAIC) Database	clinical	Legal agreement
Breast Cancer Data Set	clinical	Open
Breast Cancer Coimbra Data Set	clinical	Open
Belgian Cancer Registry (BCR)	clinical	Paid
EHR Patient's History in a Brazilian Cancer Center	clinical	Open
canSAR	clinical	Free for science
BloodPAC	clinical	Legal agreement
Liquid Biopsy Evaluation and Repository Development at Princess Margaret (LIBERATE) - Clinical study ongoing	clinical	Open
CancerID	clinical	
Twitter Triple Corpus	non clinical	Open
Switchboard Dialogue Act Corpus (SwDA)	non clinical	Open
Switchboard-1 Release 2	non clinical	Open
DailyDialog	non clinical	Open
Ubuntu Dialog Corpus	non clinical	Open
Eva courpus	non clinical	Open
BIG BAD NLP Dataset	non clinical	Open
EmpatheticDialogues	non clinical	Open
Multimodal EmotionLines Dataset (MELD)	non clinical	Open
Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)	non clinical	Open
Cornell EDU	non clinical	Open
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	non clinical	Open
Depression	non clinical	GPL 2
Face Dataset with Age, Emotion, Ethnicity	non clinical	Open
Remote collaborative and affecrive interactions (RECOLA) database	non clinical	Free for academics

The following chapters chapters will provide more information on each of the listed data-repositories and a short description of the proviced data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Clinical data repositories

This chapter lists all main repositories with clinical data by providing a short description of the content, a website link and additional information like access descriptions.

Preliminary purpose analysis:

Clinical data repositories identified here aim to expand the spectrum of available clinical data sources to identify relevant features for identification of patient cohorts and disease trajectories. Increasing the number of data qualitities and types will assist in depicting reality to a more granular extent.

The specific role and relevance of each data repository for exploitation can only be finally assessed after analysis of the data received by clinical consortial partners.

1. National Ambulatory Medical Care Survey (NAMCS)/ National Hospital Ambulatory Medical Care Survey (NHAMCS)

Physician services and hospital outpatient and emergency department services, the conditions most often treated, and the diagnostic and therapeutic services rendered, including medications prescribed.

Source: CDC Centers for Disease Control and Prevention. Department of Health & Human Services. USA

Website: https://www.cdc.gov/nchs/ahcd/index.htm

Access requirements: Open, data in the form of public health reports, journal articles, and microdata files.

Impact estimate: 8 out of 10

2. Genes-to-Systems Breast Cancer (G2SBC) Database

Integrates data about genes, transcripts and proteins reported in literature as altered in breast cancer cells.

Source: National Research Council – Institute for Biomedical Technologies. Italy.

Website: https://www.itb.cnr.it/breastcancer/

Access requirements: Open, unified query schema using a web interface, G2SBC Database relies on a MySQL server

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Impact estimate: 8 out of 10

Note: The service is currently unavailable due to maintenence.

3. National Statistics Institute (INE)

Information regarding the number of registered professionals and employment in the health and social service related sectors that have drawn a great deal of attention during this time. In addition, the data provided regarding hospital activity and deaths by cause of death facilitate the analysis, and allow comparisons to be made with the current situation and the figures currently in use to be understood. In addition to the tables selected in this section, the INE has made available a broad mining and anonymised microdata that allow for new analyses of these phenomena.

Source: National Statistics Institute INE, Spain.

Website: https://www.ine.es/covid/covid_salud.htm

Access requirements: Open

Impact estimate: 7 out of 10

4. Registry of Specialized Health Activity (RAE-CMBD)

The Registry of Specialized Health Activity (RAE-CMBD) integrates administrative and clinical information of the patients treated in different specialized care modalities, giving continuity to the CMBD but expanding, since 2016, its coverage to ambulatory care modalities and the private sector. The rules for the registration and sending of data are established in Royal Decree 69/2015 that creates the RAE-CMBD.

The statistical exploitation of these data is included among the operations of the National Statistical Plan. The consultation variables include the basic variables on the age and sex of the patient, the episode of care (discharge, intervention, visit) and the clinical variables on diagnoses and procedures to which variables derived from the use of patient classification systems and estimations of cost are added. The information is oriented to the knowledge of the demand and morbidity attended as well as to the operation and process of attention of the Specialized Care in Spain. Said information (basic data and calculated indicators) can be disaggregated and filtered by the different variables included with the only limitations that derive from the confidentiality of the data in guarantee of the personal data protection regulations.

Source: Ministry of Health and Social Wellbeing, Spain.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Website: https://pestadistico.inteligenciadegestion.mscbs.es/publicoSNS/Comun/ArbolNoArb.aspx?idNodo=23525

Access requirements: Open

Impact estimate: 6 out of 10

5. The Depresjon Dataset

The dataset was originally collected for the study of motor activity in schizophrenia and major depression. The dataset contains the following: Two folders, whereas one contains the data for the controls and one for the condition group. For each patient we provide a csv file containing the actigraph data collected over time. The columns are: timestamp (one minute intervals), date (date of measurement), activity (activity measurement from the actigraph watch). In addition, we also provide the MADRS scores in the file \emph{scores.csv}. It contains the following columns; number (patient identifier), days (number of days of measurements), gender (1 or 2 for female or male), age (age in age groups), afftype (1: bipolar II, 2: unipolar depressive, 3: bipolar I), melanch (1: melancholia, 2: no melancholia), inpatient (1: inpatient, 2: outpatient), edu (education grouped in years), marriage (1: married or cohabiting, 2: single), work (1: working or studying, 2: unemployed/sick leave/pension), madrs1 (MADRS score when measurement started), madrs2 (MADRS when measurement stopped).

Source: Simula Research Laboratory, Norway. Citation: Enrique Garcia-Ceja, Michael Riegler, Petter Jakobsen, Jim Tørresen, Tine Nordgreen, Ketil J. Oedegaard, Ole Bernt Fasmer, Depresjon: A Motor Activity Database of Depression Episodes in Unipolar and Bipolar Patients, In MMSys'18 Proceedings of the 9th ACM on Multimedia Systems Conference, Amsterdam, The Netherlands, June 12 - 15, 2018.

Website: https://datasets.simula.no/depresjon/

Access requirements: Open, reference to study required

Impact estimate: 5 out of 10

6. Medical Information Mart for Intensive Care III (MIMIC-III)

The dataset comprises 61,532 intensive care unit stays: 53,432 stays for adult patients and 8,100 for neonatal patients. The data spans June 2001 - October 2012. The database, although de-identified, still contains detailed information regarding the clinical care of patients, so must be treated with appropriate care and respect.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Source: MIT Lab for Computational Physiology. USA. Citaiton: Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35

Website: https://mimic.physionet.org/

Access requirements: Freely accessible. Researchers seeking to use the database must formally request access

Impact estimate: 7 out of 10

7. MTSamples

MTSamples contains sample transcription reports for many specialties and different work types. These sample reports are provided by various transcriptionists and users and are for reference purpose only. Medical transcription samples and reports for the following types and specialties are available on MTSamples.com:

- Allergy / Immunology - (7)
- Autopsy - (8)
- Bariatrics - (18)
- Cardiovascular / Pulmonary - (372)
- Chiropractic - (14)
- Consult - History and Phy. - (516)
- Cosmetic / Plastic Surgery - (27)
- Dentistry - (27)
- Dermatology - (29)
- Diets and Nutritions - (10)
- Discharge Summary - (108)
- Emergency Room Reports - (75)
- Endocrinology - (19)
- ENT - Otolaryngology - (98)
- Gastroenterology - (230)
- General Medicine - (259)
- Hematology - Oncology - (90)
- Hospice - Palliative Care - (6)
- IME-QME-Work Comp etc. - (16)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

- Lab Medicine - Pathology - (8)
- Letters - (23)
- Nephrology - (81)
- Neurology - (223)
- Neurosurgery - (94)
- Obstetrics / Gynecology - (160)
- Office Notes - (51)
- Ophthalmology - (83)
- Orthopedic - (355)
- Pain Management - (62)
- Pediatrics - Neonatal - (70)
- Physical Medicine - Rehab - (21)
- Podiatry - (47)
- Psychiatry / Psychology - (53)
- Radiology - (273)
- Rheumatology - (10)
- Sleep Medicine - (20)
- SOAP / Chart / Progress Notes - (166)
- Speech - Language - (9)
- Surgery - (1103)
- Urology - (158)

Source: MTHelpLine.

Website: https://www.mtsamples.com/

Access requirements: Open. Their requirement is that while linking or sharing or printing, notify them, and please give credit to their web site

Impact estimate: 3 out of 10

8. Medical Records 10 yrs

This dataset contains Lab results, diagnostic, medication details of a large number of patients over a course of 10 yrs. Language is English and data is provided in standardized csv files.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Source: Arvind Natarajan, data.world.

Website: https://data.world/arvin6/medical-records-10-yrs

Access requirements: Pricing - 1. Enterprise: several plans from 12-150$/mo, 2. Professional: $12/mo, 3. Community: free

Impact estimate: 8 out of 10

9. Distress Analysis Interview Corpus (DAIC) database

This database is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) (Gratch et al.,2014), that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder.

Source: University of Southern California, Institute for Creative Technologies. Citation: Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum DR. The Distress Analysis Interview Corpus of human and computer interviews. InLREC 2014 May (pp. 3123-3128)

Website: http://dcapswoz.ict.usc.edu/

Access requirements: Signment of aggreement required, only for academic and non-profit researchers

Impact estimate: 2 out of 10

10. Breast Cancer Data Set

This is a classic dataset used in the machine learning community for classification of cancer patients using multivariate characteristics. Data was provided by the Oncology Institute in Lubljana, and has repeatedly appeared in the machine learning literature. This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Source: UCI MachineLearning Repository. Center for Machine Learning and Intelligent Systems. USA. Creators: M. Zwitter, M Soklic. Institute of Oncology, University Medical Center, Ljubljana.

Website: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Access requirements: Open

Impact estimate: 5 out of 10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

11. Breast Cancer Coimbra Data Set

Clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls. Dataset used for multivariate classification through machine learning. There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

Source: UCI MachineLearning Repository. Center for Machine Learning and Intelligent Systems. USA. Reference: Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer, 18(1)

Website: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Access requirements: Open

Impact estimate: 3 out of 10

12. EHR Patient's History in a Brazilian Cancer Center

Sample data of patient's history in a Brazilian cancer center. A publicly available sample of one of the biggest Cancer Hospitals in southern Brazil with 13,652 samples.

Categories included in the dataset are:

- sexo (sex [0 - Male; 1 - Female])
- UTI (UCI [0 - non UCI; 1 - passed through UCI])
- freq_cardiaca (Heart Rate)
- freq_respiratoria (Respiratory frequency)
- glicemia_capilar (Capillary blood glucose)
- pa_diastolica (Diastolic blood pressure)
- pa_sistolica (Systolic blood pressure)
- sat_o2 (O2 saturation)
- temperatura (Temperature)
- outcome (0 - alive; 1 - death)

Source: Jhonatan Kobylarz

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Website: https://www.kaggle.com/jhonatankobylarz/one-of-the-biggest-brazilian-cancers-center

Acess requirements: Open

Impact estimate: 6 out of 10

13. canSAR

Integrated knowledge-base that brings together multidisciplinary data across biology, chemistry, pharmacology, structural biology, cellular networks and clinical annotations, and applies machine learning approaches to provide drug-discovery useful predictions

Source: canSAR Team | The Department of Data Science; cancer.gov, ChEMBL, clinicaltrials.gov, DepMap, DepMap (CCLE), DEPOD, GDC/TCGA, HuRI. HuRI-union, IMEx, MsigDB, Other trusted sources and publications, Pathway Commons, PDB, Phosphosite, Reactome-FI, TCGA, Tfacts. TTRUST, UniProt

Website: https://cansarblack.icr.ac.uk/#splash-about

Access requirements: freely available for use by any individual and for any purpose. canSAR is intended for use for scientific research only and cannot be used for the provision of medical advice

Impact estimate: 3 out of 10

14. BloodPAC

The Blood Profiling Atlas in Cancer (BloodPAC) Consortium was launched on October 17, 2016 to accelerate the development and validation of liquid biopsy assays to improve the outcomes of patients with cancer.

Source: BloodPAC is a consortium managed by the Center for Computational Science Research, Inc., Illinois

Website: https://www.bloodpac.org/

Access requirements: under construction repository needs legal agreement to access data common

Impact estimate: 3 out of 10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

15. Liquid Biopsy Evaluation and Repository Development at Princess Margaret (LIBERATE)

Clinical study to "incorporating the collection of peripheral blood samples (liquid biopsies) into research protocols as a means of non-invasively assessing tumor progression and response to treatment at multiple time points during patients course of disease." (NCT03702309 - ends 2022)

Cancer types included are:

- Breast Cancer
- Lung Cancer
- Colon Cancer
- Ovarian Cancer
- Melanoma
- Lymphoma
- Leukemia
- Mutation
- Lynch Syndrome
- Cowden Syndrome
- BRCA1 Mutation
- BRCA2 Mutation
- Uterine Cancer
- Myeloma
- Kidney Cancer
- Head and Neck Cancer
- Meningioma

Source: University Health Network, Toronto

Website: https://clinicaltrials.gov/ct2/show/NCT03702309

Access requirements: Individual agreement required to enter raw study data after study conclusion 2022. Aggregated data will be open source.

Impact estimate: 3 out of 10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

16. CancerID

CANCER-ID is a newly formed European consortium supported by Europe's Innovative Medicines Initiative (IMI)... thus providing a unique setting for establishing clinical utility of liquid biopsies.

Website: https://www.cancer-id.eu/news/scientific-publications/

Access requirements: Open

Impact estimate: 2 out of 10

17. Breast Cancer Data Set

This is a classic dataset used in the machine learning community for classification of cancer patients using multivariate characteristics. Data was provided by the Oncology Institute in Lubljana, and has repeatedly appeared in the machine learning literature. This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Source: UCI MachineLearning Repository. Center for Machine Learning and Intelligent Systems. USA. Creators: M. Zwitter

Website: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Access requirements: Open

Impact estimate: 4 out of 10

18. Belgian Cancer Registry (BCR)

BCR has validated cancer registration information available at the Belgian level from 2004 onwards (and currently up till 2017). For these patients, patient and tumor characteristics are available. Part of this information is directly derived from the notifications received from oncological care programs and laboratories for pathology. In addition, some information is derived from administrative data on reimbursements made by the insurance companies (provided to BCR by the InterMutualisticAgency; IMA).

Patient characteristics:

- Sex
- Age

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

- Region (at time of diagnosis)
- WHO performance status
- Comorbidities: estimated from IMA data based on medication usage in the year prior to cancer diagnosis: available for diabetes, cardiovascular and respiratory comorbidities (is not actively registered, needs additional data handling)
- Hospital admission days: count of hospitalization days in the year prior to the diagnosis with exclusion of the month before diagnosis, usually categorized. Based on IMA data. (not actively registered, needs additional data handling)
- Vital status
- Cause of death

Tumor characteristics:

- ICD-10
- TNM: clinical and pathological stage
- Differentiation grade
- Morphology
- Localization
- Multiple tumors: an indication on whether the patient has been diagnosed with other cancers than the cancer under study
- Molecular characteristics (eg ER, PR, HER2) are not actively registered but derived from pathology reports by means of text recognition tools. This data extraction will be finished for ER,PR and HER2 by autumn 2020.

Basis epidemiology reports on different cancer types including breast and colon can be found on the website: https://kankerregister.org/Cancer_Fact_Sheets.

As for the treatments, information is derived from IMA data as well, and used on a project-specific base. For breast cancer, it is regularly calculated quality of care indicators covering the diagnostic and therapeutic care path. The QCI are calculated at the individual hospital level, and feedback on the results is provided to each Belgian center. Complete QCI reports are not publicly available, but could however get a glance of some incorporated elements on this website: https://www.zorgkwaliteit.be/.

Recent publication of the Registry on cancer in elderly containing a chapter on colorectal cancers including information on treatments: https://kankerregister.org/media/docs/SKR_publicatie2018_CancerinanAgeingPopulation(landscape)5_12_2018.pdf

Website: https://kankerregister.org

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Access criteria: Reimbursement calculated on individual basis. Process for access takes several months and includes evaluation of data analysis purpose and cybersecurity analysis.

Impact estimate: 7 out of 10

Non-clinical data repositories

This capture lists all main repositories with non-clinical data by providing a short description of the content, a website link and additional information like access descriptions.

Preliminary purpose analysis:

Non-clinical data repositories identified here aim to expand the spectrum of available data sources to build the base for sensory network sensors. Hereby identified data resources are forming the backbone of the conversational intelligence developed in PERSIST

The specific role and relevance of each data repository for exploitation will be finally assessed after analysis of the data collected in studies including patient and expert surveys to identify topics and information areas most relevant to cancer survivors. These studies are conducted within PERSIST and will give insight into the focus to be chosen for the design of the mHealth app and sensory network.

19. Twitter Triple Corpus

Tweets from the sample Public stream using Twitter's streaming API, and writes them to stdout, which you can redirect to a file for later use as a corpus. This module consumes tweets from the sample Public stream and putes them on a queue. The tweets are then consumed from the queue by writing them to a file in JSON format as sent by twitter, with one tweet per line. This file can then be processed and filtered as necessary to create a corpus of tweets for use with Machine Learning, Natural Language Processing, and other Human-Centered Computing applications.

Source: bwbaugh (Github)

Website: https://github.com/bwbaugh/twitter-corpus

Access requirements: Open

Impact estimate: 6 out of 10

20. Switchboard Dialogue Act Corpus (SwDA)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Switchboard is a collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (1) no two speakers would converse together more than once and (2) no one spoke more than once on a given topic.

The Switchboard Dialogue Act Corpus (SwDA) extends the Switchboard-1 corpus with tags from the SWBD-DAMSL tagset, which is an augmentation to the Discourse Annotation and Markup System of Labeling (DAMSL) tagset. The tags summarize syntactic, semantic, and pragmatic information about the associated turn. The SwDA is not inherently linked to the Penn Treebank 3 parses of Switchboard, and it is far from straightforward to align the two resources. In addition, the SwDA is not distributed with the Switchboard's tables of metadata about the conversations and their participants. This project includes a version of the corpus (swda.zip) that pools all of this information to the best of my ability. In addition, it includes Python classes that should make it easy to work with this merged resource. This project was originally part of my LSA Linguistic Institute 2011 course Computational Pragmatics.

Source: University of Colorado at Boulder. Citation: Jurafsky, Daniel and Shriberg, Elizabeth and Biasca, Debra. University of Colorado, Boulder Institute of Cognitive Science. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. 1997,

Website: https://web.stanford.edu/~jurafsky/swb1_dialogact_annot.tar.gz

Access requirements: Open

Impact estimate: 5 out of 10

21. Switchboard-1 Release 2

Switchboard is a collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (1) no two speakers would converse together more than once and (2) no one spoke more than once on a given topic.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

The Switchboard-1 Telephone Speech Corpus (LDC97S62) consists of approximately 260 hours of speech and was originally collected by Texas Instruments in 1990-1, under DARPA sponsorship.

Source: Linguistic Data Consortium, Univeryity of Pennsylvania. Citation: Godfrey, John J., and Edward Holliman. Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium, 1993

Website: https://catalog.ldc.upenn.edu/LDC97S62

Access requirements: Open

Impact estimate: 5 out of 10

22. DailyDialog

DailyDialog: A Manually Labelled Multi-turn Dialogue.

DailyDialog is a high-quality multi-turn dialog dataset, which is intriguing in several aspects. The language is human-written and less noisy. The dialogues in the dataset reflect daily communication and cover various topics about daily life. The dataset is manually labeled with communication intention and emotion information.

The developed DailyDialog dataset contains 13,118 multi-turn dialogues

Total Dialogues	13,118
Average Speaker Turns Per Dialogue	7.9
Average Tokens Per Dialogue	114.7
Average Tokens Per Utterance	14.6

Source: PolyU Hong Kong. Citation: Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. IJCNLP 2017.

Website: https://github.com/Sanghoon94/DailyDialogue-Parser

Access requirements: Open. Only for research purposes.

Impact estimate: 5 out of 10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

23. Ubuntu Dialog Corpus

The Ubuntu Dialogue Corpus consists of almost one million two-person conversations extracted from the Ubuntu chat logs, used to receive technical support for various Ubuntu-related problems. The conversations have an average of 8 turns each, with a minimum of 3 turns. All conversations are carried out in text form (not audio).

The full dataset contains 930,000 dialogues and over 100,000,000 words and is available here. This dataset contains a sample of this dataset spread across .csv files. This dataset contains more than 269 million words of text, spread out over 26 million turns.

Source: McGill University, Univerité de Montreal, Canada. Citation: Ryan Lowe, Nissan Pow, Iulian V. Serban and Joelle Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems", SIGDial 2015. URL: http://www.sigdial.org/workshops/conference16/proceedings/pdf/SIGDIAL40.pdf

Website: https://www.kaggle.com/rtatman/ubuntu-dialogue-corpus

Access requirements: Open

Impact estimate: 5 out of 10

24. Eva courpus

The EVA Corpus consists of four audio/video sessions plus corresponding orthographic transcriptions, each with a duration of 57 minutes, and 228 minutes in total. The multi-party spontaneous discourse in all four sessions is from an entertaining evening TV-talk show, As ti tud not padu?!, broadcast by the Slovene commercial television (PoPTV). The show contains casual conversation about general, informal and personal aspects of interviewee's life

The transcribed session of this recording has been annotated using ELAN 4.9.4. In addition to the original transcription and morphosyntactic annotation from the GOS corpus, the following layers of information are added:

- statement sentiment
- phrase breaks within statements
- prominence of statements
- sentences within the statement
- sentence sentiment
- sentence type

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

- speaker visibility on the scene
- gesture units
- gesture phrases
- emotions
- semiotic intent
- dialogue role

Source: University of Maribor, Slovenia. Citation: Mlakar, Izidor; Majhenič, Simona; Rojc, Matej and Verdonik, Darinka, 2020, Multimodal corpus EVA 1.0, Slovenian language resource repository CLARIN.SI.

Website: https://www.clarin.si/repository/xmlui/handle/11356/1311

Access requirements: Open

Impact estimate: 6 out of 10

25. BIG BAD NLP Dataset

The Big Bad NLP Database is the world's largest data library in natural language processing: including 481 datasets including datasets in +36 different languages provided in common data formats (JSON, xml, CSV, etc) covering various domains in NLP. Besides the inclusion of classic datasets found in GLUE and SuperGLUE, we also have included datasets ranging from the humongous CommonCrawl to the classic Penn Treebank.

It does not solely focus on classic NLP tasks either. While including standards for classifying, question answering, we've also covered datasets pertaining to text-to-SQL, speech recognition, and multi-modal (text and images).

Website: https://datasets.quantumstat.com/

Access requirements: Open

Impact estimate: 9 out of 10

26. Depression

The dataset is involved into the analysis of depression. The data consists of a study about the life conditions of people who live in rurales zones. The dataset has 23 columns and a total of 1432 registered objects.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

The features or columns included are the following:

- Survey id Ville id
- Sex
- Age
- Married
- Number children educationlevel
- Total members (in the family) gained asset
- durable asset save asset
- living expenses other expenses
- incoming salary incoming own farm incoming business
- incoming no business
- incoming agricultural farm expenses
- labor primary lasting investment
- no lasting investmen
- depressed: [Zero: No depressed] or [One: depressed] (Binary for target class)

Source: Deigo Babativa, Kaggle.com.

Website: https://www.kaggle.com/diegobabativa/depression

Access requirements: Open source (GPL 2)

Impact estimate: 3 out of 10

27. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

Source: S. Livingstone. Ryerson University, Toronto, Canada. Citaiton: Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Website: https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

Access requirements: Open

Impact estimate: 7 out of 10

28. Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).

Source: Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE transactions on affective computing. 2014;5(4):377-390. doi:10.1109/TAFFC.2014.2336244.

Website: https://github.com/CheyneyComputerScience/CREMA-D

Access requirements: Open (Open Database License)

Impact estimate: 6 out of 10

29. EmpatheticDialogues

Provides a novel dataset of 25k conversations grounded in emotional situations. Non-clinical

Source: Facebook AI Research. Citation: H. Rashkin, E. M. Smith, M. Li, Y. Boureau Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset

Website: https://github.com/facebookresearch/EmpatheticDialogues

Access requirements: Open source (Creative Commons)

Impact estimate: 8 out of 10

30. Multimodal EmotionLines Dataset (MELD)

A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations is a non-clinical dataset for Emotion Recognition in Multiparty Conversations. Multimodal EmotionLines Dataset (MELD) has been created by enhancing and extending EmotionLines dataset. MELD contains the same dialogue instances available in EmotionLines, but it also

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

encompasses audio and visual modality along with text. MELD has more than 1400 dialogues and 13000 utterances from Friends TV series. Multiple speakers participated in the dialogues. Each utterance in a dialogue has been labeled by any of these seven emotions -- Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear. MELD also has sentiment (positive, negative and neutral) annotation for each utterance.

Source: SUTD, Singapore, National University of Singapore, Singapore, Instituto Politecnico Nacional, Mexico, Nanyang Technological University, Singapore, University of Michigan, USA. Citation: S. Poria, D. Hazarika, N. Majumder, G. Naik, R. Mihalcea, E. Cambria. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation. (2018)

Website: https://affective-meld.github.io/

Access requirements: Open

Impact estimate: 6 out of 10

31. Cornell EDU

Extensive collection of NLP sources curated and provided by the Cornell University, USA.

Datasets included are:

- Argument trees, "successful persuasion" metadata, and related data from the subreddit ChangeMyView (first release 2016; 321MB)
- Multimodal posts for popularity prediction. (first release 2017; 3.3GB)
- 1B Comments and posts for highly-related ("affix pair") subreddits. The starting point was data collected and released by Jason Baumgartner with additional processing done for the dataset below. (first release 2016; 13 GB)
- Multi-community engagement (users posting, or not posting, in different subreddits since Reddit's inception). Data includes the texts of 76.7M posts made and associated metadata, such as the subreddit, the "number" of upvotes, and the time stamp. The starting point was data collected and released by Jason Baumgartner. (first release 2015; 24GB)
- Multimodal datasets for quantifying visual concreteness (first release 2018; Wikipedia dataset: 4.9GB; British Library dataset: 38GB)
- 1977-2008 FOMC transcripts: multiple many-hour meetings where very consequential decisions (what the US Federal Interest Rate will be) are made between participants who know each other very well. (First release 2016; 231M when unzipped)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

- Cornell natural-experiment tweet pairs: data for investigating whether whether phrasing affects message propagation, controlling for user and topic. Note that, in compliance with Twitter policy, we cannot distribute the tweets themselves, but rather tweet IDs. zip file can be retrieved from the given URL (first release 2014)
- Cornell movie-quotes corpus: paired memorable and non-memorable movie quotes, controlling for speaker, scene, and length (first release 2012)
- Supreme Court dialogs corpus: conversations and metadata (such as vote outcomes) from oral arguments before the US Supreme Court (first release 2012)
- Wikipedia editor conversations corpus (first release 2012)
- GMOHedging: data for studying hedging and framing in GMO debates and in professional- vs. pop-science discourse (first release 2012)
- Cornell movie-dialogs corpus: conversations and metadata (IMDB rating, genre, character gender, etc.) from movie scripts (first release 2011)
- Files associated with extracting lexical-level simplifications from Simple Wikipedia (first release 2010)
- Cornell movie-review corpus: Sentiment-classified movie reviews (positive/negative or number of stars), subjective/objective sentences, etc. (released in 2002/2004/2005)
- Convote: Congressional floor-debate transcripts, with support/oppose labels (first release 2006)
- Search-set results for review-oriented queries, with subjective/objective labels (first release 2008)
- AP88 data for some similarity-based pseudoword disambiguation experiments
- Multi-parallel proof/verbalization data for a project on verbalizing NuPrl mathematical proofs using multiple-sequence alignment

Source: Cornell University, USA.

Website: http://www.cs.cornell.edu/home/llee/data/

Access requirements: Open

Impact estimate: 5 out of 10

32. Face Dataset with Age, Emotion, Ethnicity

An image bounding box dataset with faces annotated to features.

The images are annotated with 4 different labels:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

- Age Range
- Ethnicity
- Gender
- Face emotion

Source: DataTurks. Kaggle.com

Website: https://www.kaggle.com/dataturks/face-dataset-with-age-emotion-ethnicity

Access requirements: Open source

Impact estimate: 5 out of 10

33. Remote collaborative and affective interactions (RECOLA) database

The RECOLA database consists of 9.5 hours of audio, visual, and physiological (electrocardiogram, and electrodermal activity) recordings of online dyadic interactions between 46 French speaking participants, who were solving a task in collaboration. Affective and social behaviors naturally expressed by the participants were reported by themselves, at different steps of the study, and by six French-speaking assistants using the ANNEMO web-based annotation tool (time and value 'continuous'), for the first five minutes of interaction; 3.8/2.9 hours of annotated audiovisual/multimodal data, respectively.

Website: https://diuf.unifr.ch/main/diva/recola/download.html

Access requirements: User account required, free for academics and non-profit organisations (EULA), commercial license is available.

Impact estimate: 3 out of 10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875406

Conclusion

In conclusion the list of additional data repositories diligently prepared as part of workpackage 2 will be a solid preparation for workpackages 4 and 5, which will help to mitigate potential risks harboured in the nature and quality of data provided by clinical partners of the consortium to produce novel insights into patient trajectories and intervention points. Therefore maximizing results of the big data platform and artificial intelligence algorithms.

As described earlier, clinical data repositories identified here aim to expand the spectrum of available clinical data sources to identify relevant features for identification of patient cohorts and disease trajectories. Increasing the number of data qualities and types will assist in depicting reality to a more granular extent. Non-clinical data repositories will build the base for sensory network sensors.

The specific role and relevance of each data repository for exploitation can only be finally assessed after analysis of the data received by clinical consortial partners on the one hand and on the other hand after analysis of the data collected in studies including patient and expert surveys to identify topics and information areas most relevant to cancer survivors. These studies will be conducted within PERSIST in workpackage 8 and will give insight into the focus to be chosen for the design of the mHealth app and sensory network.