Project # 643735
www.do-change.eu

Data Analysis and Big Data Analytics Phase 2 - Design

Report 4.18-D69 Draft Version 0.2

Key Information from the DoA	
Due Date	M38
Type	OTH
Security	Public
Description:	

This document reports the Data Analysis and Big Data Analytics Design for Phase II of the Do-CHANGE project.

Lead Editor: G. Gavidia (EUT) **Internal Reviewer:** M. Wetzels (TUE)

Versioning and contribution history

Version	Date	Author	Partner	Description
Draft 0.1	19-Jun 2018		EUT	1st Draft Version
Draft 0.2	20-Jun-2018		EUT	2nd Draft Version
Draft 0.3	22-Jun-2018		EUT	Includes corrections provided by M. Wetzels

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

- ***Data Sources***

  Data gathered in phase 2 is longitudinal, i.e. several outcome variables have been measured repeatedly for the same participant along the day at multiple days.

  - "DOs" are messages that are intended to stimulate behaviour change. DOs – textual messages, optionally with an associated URL providing further information, which is intended to promote a lifestyle change.
  - Clinical Data: this dataset includes data from self-reported questionnaires. Data gathered involves information about blood pressure, ECG, symptoms, medication, alcohol consumption and smoking.
  - Physical Activity Data: data gathered from activity trackers and mobile applications including Fitbit, Beddit and Moves.
  - Psychological Data, which included information from several self-reported questionnaires focused on different individual's aspects such as:
    - Emotional states: Patient Health Questionnaire (PHQ-9), Generalised Anxiety Disorder Questionnaire (GAD-7) and Standard Assessment of Negative Affectivity, Social Inhibition, and Type D Personality
    - Lifestyle: Health Promoting Lifestyle Profile (HPLP-II)
    - Behavioural flexibility: Do Something Different
    - Quality of Life: WHO Quality of Life questionnaire (WHOQOL-bref)

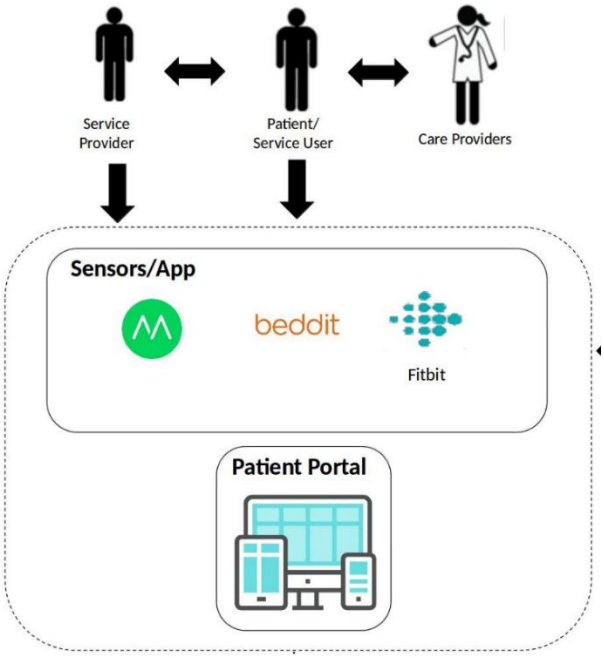

Figure 1: Data Sources

- ***Pre-processing Stage***
  - ***Data cleaning and filtering*** methods are applied to ensure the overall consistency of the data and the integrity of results by detecting and correcting (or removing) corrupt and inaccurate values.
  - ***Feature construction***: this step is focused on converting "raw" data into a set of useful features. Here, additional variables are built from pre-existing variables in order to obtain more informative and non-redundant data. For example, by using age information, participants are stratified by age groups including aged<50 years, aged [50-65) and aged≥65 adults.
  - ***Feature Selection***: consists in selecting a subset of relevant and informative variables to be used as input variables in the model building stage. The main reasons to apply feature selection are: (1) simplify the models by making their training faster and being easier to interpret, (2) improve the accuracy of models and (3) reduce overfitting (i.e. improve generalisation capabilities).
  - ***Down Sampling***: the pre-processing stage includes down sampling the data for a given time interval length by adding up the steps every week.
- ***Data Analysis***
  - ***Relationship***
    - **Descriptive Statistics**. Descriptive statistics are computed for baseline observations in order to evaluate the characteristics and significance of variables across groups of stratification such as diagnostic groups and cultural groups. Significant differences were evaluated at P value < 0.05.
    - **Correlation and Mutual Information**: both methods are applied to determine the relationship between variables pairs. Correlation is applied to continuous variables, whereas Mutual Information is applied to categorical variables.
  - ***Profiling:*** participant and population profiling is carried out to study the temporal pattern of the target variables.
  - ***Change Detection:*** this stage is focused on detecting and analysing human behaviour change.
    - **Activity density maps**: track different types of changes in physical activity data, time periods or windows based on weeks were quantitatively and objectively compared by using activity density maps. This way, if two time windows contain significantly different physical activity data; then this may indicate a significant behaviour change.
    - **Comparison of means** is applied to compare the change of continuous variables over time, mainly on a weekly basis. This technique allows to determine whether there is a significant change by comparing follow-up observation to baseline observations.
    - **Linear Mixed Effect Modelling (LME)**: longitudinal data have two sources of variability: (1) inherent within-subject change, (2) between-subject variation. The inherent within-individual variation is a consequence of some subject-specific biological process that progresses gradually over time. The second source of variability reflects natural variation in the individual's measurement trajectory. In this part, we aimed to estimate the change of our main research variables over time as well as to assess group differences between the trajectories of

variables of interest. For that purpose, a suitable method for longitudinal data analysis called LME is applied. LME takes into account the within and between individual variability. LME modelling also takes into account the group level structure in the data by simultaneously assessing effects within and across groups. LME models incorporate both fixed-effects and random-effects, and; describe the relationship between a response and the covariates that have been observed along with the response.

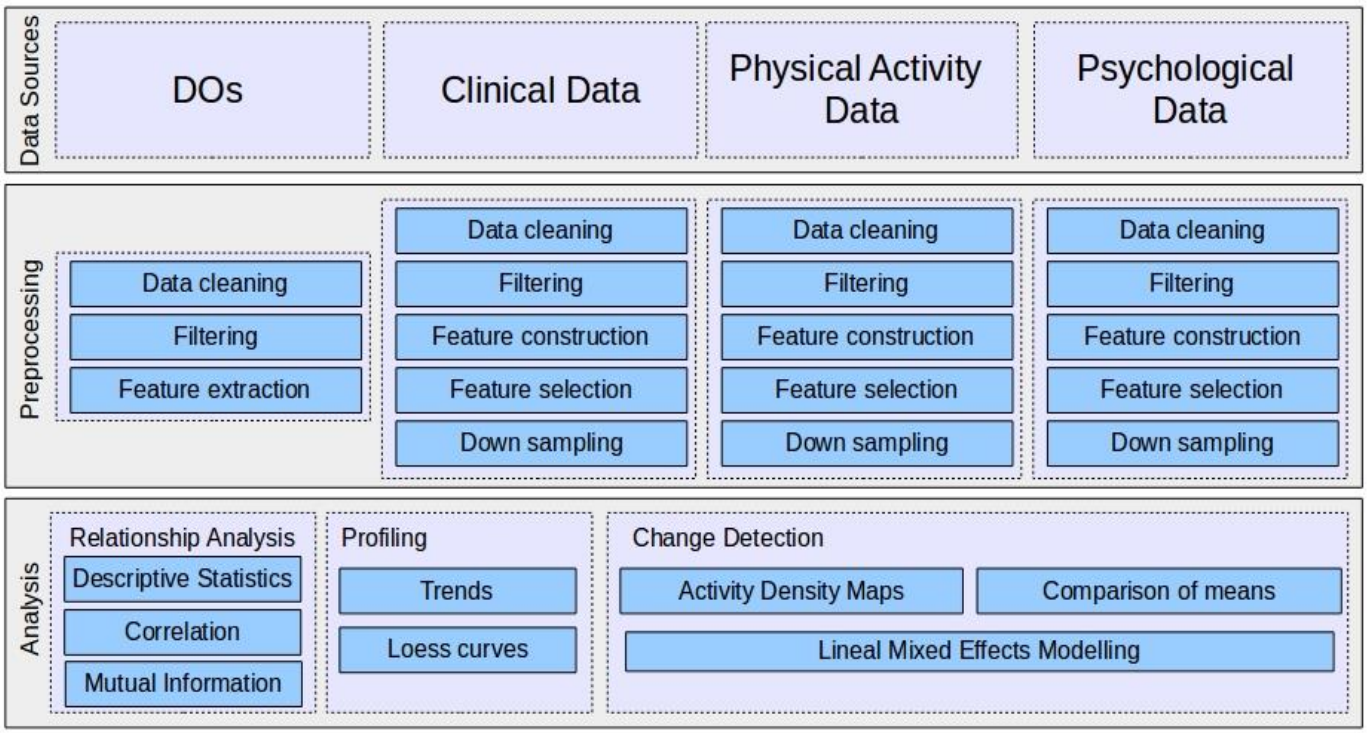

Figure 2: Data Analysis Framework