README

This folder is made up of five files including:

- README_challenge2.pdf: the present readme file
- data_challenge2.xlsx : an excel file containing the measurements on 928 objects (patients) in rows and 121 variables in columns (plus an additional id variable in the first column). The first row of entries gives the variable names.
- variables_data_challenge2.xlsx: an excel file giving further details of the variables' descriptions (in corresponding order to their appearance in the dataset file) as well as details of data types, missingness, any preprocessing that has taken place, as well as references.
- questionnaire_data_challenge2.pdf: a completed version of the standard questionnaire for data in the benchmarking archive including meta-data details of the source of the data, scientific background, cluster goals and their justifications, desired cluster characteristics and their justifications
- call_challenge2.pdf: a document detailing the challenge rules and deadline.

Inputs for the clustering: 112 variables from bsex0 (11th variable) to Start_risk (122nd variable). Not all variables need to be included in the analysis. Variables 2 to 10 should not be used for the actual clustering but could be used to inform or validate the clustering.

Metadata: this information is available in the questionnaire_data_challenge2.pdf file (particularly the Quality Concerns and Justification of Clustering sections) indicating desired characteristics of the resulting clustering.

Outputs: These should include (at a minimum):

- A short report detailing the analyses applied to the data and the resulting outcome(s), including a visualization and an evaluation of the result(s), and the code used for the analysis (or, in case proprietary software was used, detailed information on the version of the software and on how the analysis with it was carried out).
- For each clustering result, a csv file with a matrix indicating cluster membership of the objects clustered. The first column of this matrix should contain the id numbers of all objects in the clustering, each subsequent column should represent one of the clusters or a noise/outlier set (if one is required). The entries in the matrix can be binary or non-binary (either cluster membership indicators or probabilities). If the clustering allows it, objects can be assigned to more than one cluster.