TITLE OF THE PROJECT: Analysis of big data in ageing using clustering and exploratory methods

HEAD OF THE TEAM: Sandrine Andrieu, Inserm UMR 1027 - Aging and Alzheimer disease: From observation to intervention

SUMMARY

Background

The CONSTANCES study will recruit approximately 66 000 individuals aged 45 and older and will collect thousands of variables from multiple sources (clinical, biological and behavioural evaluations, postal questionnaires, administrative databases), making it much larger than typical studies of ageing. It will therefore provide a unique opportunity to study subgroups of individuals, which would not be possible in typical cohort studies. However, the collection of such large datasets is associated with major financial, technical and analytical constraints. Future studies of ageing may wish to know whether or not it is worth the effort to collect so much additional data, and what extra information can be gained from this? The overall aims of this project are: (1) to identify clusters of individuals and the predictive values of these clusters, and (2) to provide bases for the design and analysis of future studies of ageing in the era of big data.

Objectives

1. In order to identify groups of individuals with similar health factors and prognosis:

    a) Identify clusters of individuals with similar characteristics across multiple domains at baseline, and describe the discriminant factors of these clusters.

    b) Study the relationship between these clusters and cognitive and functional decline in the short-term and the onset of dementia, cardiovascular disease, cancer and mortality in the long-term.

    c) Determine whether the clustering of individuals remains stable over time or whether it is influenced by the addition of follow-up data.

2. In order to inform the design and analysis of future studies of ageing with limited resources:

    a) Determine to what extent decreasing sample size and/or number of variables affects the underlying latent structure of the dataset

    b) Determine the impact of decreasing sample size, number of variables or type of statistical methods on the results of analyses aiming to identify factors associated with poor functional status at baseline or functional decline

Methods

Cross-sectional and longitudinal analyses will be conducted on data from CONSTANCES participants aged ≥45 at baseline. The analyses will be data- rather than hypothesis-driven, so all available data from the different sources will be analysed as explanatory and/or descriptive variables.

The methods used to fulfil each objective are described below:

1a. Clusters of individuals with similar characteristics across multiple domains will be identified in the baseline dataset using multidimensional data analysis methods, in particular clustering approaches (hierarchical clustering, double clustering).

1b. The relationship between baseline clusters and cognitive (MMSE) and functional (IADL) decline in the short-term (5 years), and the onset of dementia, cardiovascular disease, cancer and mortality in

the long-term (≥10 years), will be studied using methods such as mixed effects models, and survival analyses, using cluster membership as the explanatory variable.

1c. The clustering analyses described in 1a. will be repeated after the after the addition of 5-year follow-up data, and will be constrained to the same number of clusters as in 1a. Multiple Factor Analysis will be used to study the evolution of the characteristics constituting the clusters.

2a. In order to determine if the latent structure implicitly present in the full dataset remains present in samples with fewer subjects and/or variables, successive exploratory factor analyses (EFA) will be performed on various restricted samples of the dataset.

2b. Analyses aiming to identify factors associated with functional (IADL) status and decline will be performed in the full dataset and results compared with those obtained in restricted samples of the dataset with fewer subjects and/or variables. The results obtained using different methods of analysis, both traditional (e.g. logistic regression, Cox proportional hazards & mixed effects models) and non-traditional (e.g. regression trees, linear discriminant analysis), will also be compared.

Perspectives

The analyses conducted for objectives 1a – 1c could provide useful information for personalized medicine approaches, while those conducted for objectives 2a and 2b will help to determine the "cost-effectiveness" of such large-scale studies of ageing and the optimal methods of analysis for big datasets in epidemiological research on ageing.

Note: this project is part of the research consortium 'PRESAGE – PREparing Successful AGEing'