Biostatistics – Methods

Start date: March 2016 **Completion date: December 2019**

Executive Summary of Project

The use of a diverse set of clinical and biological markers, alongside genetics and other traditional risk factors can lead to more refined models for predicting an individual's risk of rapid cognitive decline or onset of dementia, increased understanding of disease processes and insight into potential new therapeutic targets. Additionally, these models can help identify individuals in the early stages of disease who are at highest risk and therefore potential candidates for recruitment into clinical trials, as well as help tailor management of subjects based on their estimated individualised risk. WP14B aimed to develop appropriate statistical methodology for dementia risk stratification and prediction using individual-level data, obtained from DPUK cohorts accessed through the DPUK portal. Our Bayesian semi-supervised mixture modelling methodology focussed on integrating data from different modalities, some of which are collected longitudinally, to identify subgroups of individuals who have particular profiles of cognitive, imaging and genetic profiles that are indicative of higher risk of disease progression. Due to various issues concerning timely access to data from the targeted DPUK cohorts through the DPUK portal and availability of genetic and biological data, through the portal, we developed our models primarily using ADNI and CFAS I. Analyses using linked ELSA cohort and ELSA METADAC genetic data obtained in 2020 are ongoing.

Team members funded (full or part-time) by DPUK

Anais Rouanet, Steven Hill, Mary Fortune

Team members involved with the project but not funded by DPUK

Brian Tom, Sylvia Richardson, Fiona Matthews and Adrian Mander

ECR's: Anais Rouanet, Mary Fortune

Locations: MRC Biostatistics Unit, University of Cambridge

Objectives

To develop and apply state-of-the-art stratification methods to DPUK cohorts. It will also provide exemplar statistical analyses to test and demonstrate the utility of the informatics portal for integrated analyses.

1. To develop robust (statistical) strategies/algorithms to identify risk stratified subgroups, which can be recruited either into Experimental Studies (including clinical trials) or followed for disease progression based on cross-sectional information.
2. To develop robust statistical strategies/algorithms for trajectory risk stratification that can identify subgroups of individuals in pre-clinical/prodromal stages of disease who have different progression rates to dementia using longitudinal information.
3. To statistically identify, incorporate and evaluate varying types and complexities of biomarkers (e.g. structural MRI, urine or blood markers) or combinations of biomarkers that associate with brain Aβ pathology or with onset of dementias or predictive of treatment response.
4. To provide model-based statistical input, driven by an iterative strategy, which can inform clinical trial designs for dementias-related experimental studies.

Dependencies to and from other work packages, networks and themes

N/A

Were all Milestones completed – No

- M1.1.3 is now met. M1.1.4 is yet to. Due to difficulties in getting access to data on a combination of biomarkers, cognitive and genetic data through the portal, we initially began working on the CFAS data which we applied for directly through the CFAS data access process, whilst awaiting for the various legal arrangements to be put in place between DPUK and the Cohorts so as to allow access to data through the DPUK informatics portal. Although CFAS I clinical data are now accessible through the Portal, the International Genomics of Alzheimer's Project (IGAP) CFAS SNP data are not available through the Portal. As CFAS I does not have any biomarker data, we concentrated our analyses on the CFAS I cognitive and IGAP SNP data. Using these CFAS I data we proposed and fitted a Profile regression model, in conjunction with a screening model to identify relevant SNPs. The outcome was a binary indicator of rapid cognitive decline between baseline and first follow-up (five years later)

and the profile variables were the identified SNPs. This work was presented at various meetings. Unfortunately, when we attempted to validate using ADNI, we found that there was poor transportability. Under further investigation, we identified that a number of the screened SNPs showed conflicting directions of univariate association with cognitive decline between CFAS I and ADNI. This thus prevented us from publishing this piece of work as we could not replicate it. We therefore turned to exploring using the same methodology with the same cognitive outcome but now including a polygenic risk score (PRS) instead of separate SNPs. The write-up of this latter work has yet to be completed as we were awaiting the ELSA METADAC genetic data in order to further test/validate the findings. These ELSA METADAC genetic data linked to the ELSA Cohort data became only recently available in 2020. As both ELSA and CFAS I did not have imaging, we could not incorporate imaging to these models using these Cohorts. We instead used ADNI, which is not a DPUK cohort. However, we will use the Lothian Birth Cohort of 1936 to do this over the next few months.

- Although a strategy for M3.1.1 is now in place, we have not been able to pursue this as yet. The methodology extends the Bayesian semi-supervised mixture models with covariates and longitudinal outcomes used in Objective 2 to further include a time to event outcome, such as time-to-onset of dementia. Unfortunately, due to Anais Rouanet leaving early to take up a post in Bordeaux in September 2019, we have yet to implement this work and apply it to an appropriate dataset. Discussion with Anais Rouanet is ongoing on how best to complete this work.

Deliverables	Milestones	Milestone deadline	Work package dependencies	Person(s) responsible
Objective 1:				
D1.1 Validated statistical model to identify risk stratified groups	M1.1.1 Access cross-sectional data	M1.1.1 Complete	To and from WP1 and WP2	Brian Tom
	M1.1.2 Develop test and validate model	M1.1.2 Complete		
	M1.1.3 Integration of imaging data into models	M1.1.3 Complete		
	M1.1.4 Paper submitted to peer review journal	M1.1.4 Dec 2019		
Objective 2:				
D2.1 Validated statistical model to identify subgroups of individuals	M2.1.1 Exploring the utility of mixed models and completion of model development	M2.1.1 Complete	Dependent on WP1 and WP2	Brian Tom
	M2.1.2: Comparison of different semi-supervised clustering approaches under classical and Bayesian frameworks (incl. using Gaussian Processes)	M2.1.2 Complete		
	M2.1.3 Paper submitted to peer review journal	M2.1.3 Complete		
Objective 3:				
D3.1 Implement a strategy for statistical identification and evaluation	M3.1.1 Strategy in place and statistical activity underway	M3.1.1 Dec 2019	None	Brian Tom
Objective 4:				
D4.1 Implement a strategy for model-based statistical input	M4.1.1 Strategy in place and statistical activity underway	M4.1.1 Complete	None	Brian Tom

Lessons Learnt

- We learnt not to assume that access to DPUK cohorts through the portal would be forthcoming. Even with contingencies in place to get access to cohort data externally than through the portal, this took an extremely long time to complete as we had to not only formally apply again by submitting applications of the project to the various Data Access Committees, go through the various datasets, select the variables required (which was extremely time-consuming), clarify the role of DPUK in the applications (in particular, with regard to obtaining ELSA genetics via METADAC), add appropriate expert on the application to deal with possible incidental findings, but after approval we then needed the University to put the data transfer agreements in place (e.g. Whitehall II) after seeking appropriate clarification and

sort out with DPUK the payments for access to some of these cohorts (such as UK Biobank and METADAC ELSA Genetics). For example, the process of getting access to the linked METADAC ELSA Genetic and Cohort data took over a year to complete.

- Interaction with other researchers in dementia research and gaining exposure of our work through meetings went well
- Skills developed using the remote secure portal and data environment
- Further lesson learnt included clarifying whether genetic data of DPUK Cohorts (if exist) would be made available through the DPUK portal or externally through a separate application, the arrangements involved with getting access to UK Biobank data and on using them on through the DPUK portal

What is the most successful outcome and what does it mean for future dementia research?

The development of methodology that integrates in a coherent way the multiple domains/dimensions of risk without having to summarise as a risk score. The model (i) allows the identification of subgroups that are linked to different risk profiles, thereby helping to identify potential individuals for recruitment into clinical trial; (ii) allows the prediction of disease course (iii) allows further personalising of management/treatment strategies; and (iv) allows improved understanding of the disease and potentially allows for the identification of biomarkers that could be targeted for therapeutic development.

Outputs

PUBLICATIONS:

Top five:

- Rouanet, A, et al. (2020). **Benefit of Bayesian clustering of longitudinal data: study of cognitive decline for precision medicine.** Book chapter in "Bayesian Methods in Pharmaceutical Research" edited by E. Lesaffre, G. Baio and B. Boulanger

https://www.crcpress.com/Bayesian-Methods-in-Pharmaceutical-Research/Lesaffre-Baio-Boulanger/p/book/9781138748484

We have developed a Bayesian Dirichlet Mixture model with Gaussian Process priors for identifying subpopulations of patients with different covariate profiles which are linked to different cognitive functioning trajectories. We identified four subpopulations with differing longitudinal cognitive trajectories linked to profiles described by 6 MRI volumetric imaging biomarkers, gender, APOE4 carrier status and educational attainment. One subpopulation is associated with steep cognitive decline and characterised by low levels of hippocampal and entorhinal cortex volume and high prevalence of APOE4 carriers and low proportion with 16 or more years of education. Persons identified as belonging to this cluster earlier on in their cognitive decline can be managed more intensively or be recruited into clinical trials. Our methodology allows us to predict future cognitive decline in subjects based on covariate profiles and cognitive functioning history.

This work has been presented (orally or as a poster) at a number of meetings including DPUK Analyst meeting, DPUK, ARUK, RSS and EcoStat conferences

COLLABORATIONS & PARTNERSHIPS:

- Investigator on Deep and Frequent Phenotyping Project
- Investigator on EPAD
- Investigator on Alzheimer's Society grant on Stratified Cohort based on Dementia Risk

FURTHER FUNDING:

- MRC grant entitled "Deep and Frequent Phenotyping: combinatorial biomarkers for dementia experimental medicine"
- Alzheimer's Society grant entitled "Bioresource - Genes and Cognition. Establishing a stratified population cohort of 100000 people recallable for pre-clinical studies of neurodegeneration and dementia"

NEXT DESTINATIONS:

- Anais Rouanet has gone to the University of Bordeaux to work with Cecile Proust-Lima and is continuing to work on dementia
- Mary Fortune is a Teaching Associate in Medical Statistics and Assessment in the Public Health Education Group

- Steven Hill is still at the MRC Biostatistics Unit

ENGAGEMENT ACTIVITIES:

2019

- B. Tom. Non-parametric clustering for longitudinal cognitive measurements, baseline imaging and genetic data. EcoSta 2019, Taichung, Taiwan 25th June 2019 (Invited Speaker)
- A. Rouanet, B. Tom, S. Richardson. Nonparametric clustering approach for longitudinal cognitive measurements, baseline imaging and genetic data in precision medicine (oral presentation), Channel Network Conference 2019, Rothamsted, UK
- A. Rouanet, S. Richardson, B. Tom. Bayesian nonparametric clustering from longitudinal cognitive measurements, baseline imaging and genetic data for precision medicine (oral presentation), ISCB 2019, Leuven, Belgium

2018

- B. Tom. Mixture models for stratifcation in dementia research. DPUK Next Generation Seminar: Analytics, Royal Statistical Society, London, UK (Invited Speaker)
- A. Rouanet, R. Johnson, S. Richardson, B. Tom. Dirichlet process mixture model for longitudinal data and side information for precision medicine: Study of cognitive decline (oral presentation), IBC 2018, Barcelona, Spain
- A. Rouanet. Bayesian outcome-driven mixture modelling of a longitudinal marker and profile variables: Precision medicine in Alzheimer's Disease (oral presentation), RSS 2018, Cardiff, Wales
- M. Fortune and A. Mander. Designing Dementia Trials Embedded within a Cohort (poster) ARUK 2018
- S. M. Hill, S. Richardson and B. D. M. Tom. Risk stratification for cognitive decline using genetic data (poster) ARUK 2018
- A. Rouanet, R. Johnson, S. Richardson, B. Tom. Identification of subgroups with specific cognitive evolution patterns and brain imaging profiles for precision medicine (poster) ARUK 2018

RESEARCH TOOLS & METHODS:

- In addition, our R software (**PReMiuMar**) which allows Bayesian clustering to be extended to longitudinal data is now available at https://github.com/anarouanet/PReMiuMar. This software implements both the multivariate normal and Gaussian Process extensions to handle longitudinal data in this framework. The accompanying paper to this R software package extension is near completion.

RESEARCH DATABASES & MODELS:

- We have undertaken an investigation into the utility of genetic markers and polygenic risk scores for rapid cognitive decline for the purpose of risk stratification. We have used IGAP genetic and cognitive data from CFAS I and corresponding data from ADNI to perform this investigation.
- We have compared latent class mixed modelling methodology to Bayesian profile regression methodology for AD research; and further investigated extensions of these methodologies for handling multivariate longitudinal outcomes and event history outcomes, for incorporating prior knowledge and for improving the efficiency and scalability of the MCMC algorithm.
- We have been involved in DPUK meetings to improve the accessibility, functionality and relevance of the DPUK Data Portal for research; providing feedback and identifying issues/hurdles.

SOFTWARE & TECHNICAL PRODUCTS:

- R software (**PReMiuMar**) which allows Bayesian clustering to be extended to longitudinal data is found at https://github.com/anarouanet/PReMiuMar.

AWARDS & RECOGNITION:

- Steven Hill has received a DPUK travel award for his presentation at the DPUK 2017 Annual Scientific Conference

WP 14b End report

Introduction

Dementia is one of the most challenging global health problems of the 21st century, affecting around 50 million people globally with numbers expected to rise substantially over the next thirty years (WHO, 2018). The repeated failure of promising Alzheimer's disease (AD) drugs to obtain regulatory approval has shifted research towards secondary prevention and clinical trials in individuals at either the pre-clinical or asymptomatic stage of disease where individuals are biomarker positive for AD. Identifying persons early in disease through use of biomarkers may increase the likelihood that treatments will be more effective in slowing or arresting further progression of the disease. However, there is currently a dearth of disease progression models capable of accurately identifying individuals at "high risk" of progression to dementia or of rapid cognitive decline. Such models have the potential to reduce screen failures and increase the chance of demonstrating effectiveness in secondary prevention trials. Moreover, use of them to identify potential biomarkers that can be surrogates of clinical outcomes in trials would be appealing, as changes in these biomarkers would be expected to be seen earlier than current clinical endpoints, such as those based on cognitive functioning (especially in pre-clinical and asymptomatic populations). This would then enable shorter duration and more cost-efficient clinical trials to be conducted. Furthermore, accurate disease progression models in dementia would be extremely useful for making more informed clinical management (and treatment) decisions and for better understanding of the aetiology of disease.

DPUK is developing an informatics portal where researchers can access data from DPUK cohorts and apply statistical methods to analyse these data. The models we develop will be used to test and demonstrate the utility of this portal.

In persons with Alzheimer's disease, there is substantial heterogeneity in disease course. For example, some individuals show rapid cognitive decline, other may stay cognitively stable for a relatively long period of time, whilst others show gradual decline. The reasons underlying these differences in clinical phenotype are not well understood currently, although it is clear that genetic, biological and environmental factors are all involved.

Stratified or precision medicine aims to (i) uncover biologically and clinically meaningful subpopulations of individuals within heterogeneous disease and clinical population; and (ii) more accurately and differentially diagnose, monitor and predict disease course, risk and response to treatment. In WP14B, we aim to develop and apply statistical methods to DPUK cohorts in order to identify subgroups of individuals that are at high risk of rapid cognitive decline or onset of dementia. However, instead of measuring risk through a single "risk" score which combines multiple domains, we propose instead to jointly model the various information from different modalities in order to better capture the complex multi-dimensional risk spectrum. That is, we develop methodology which characterises phenotypic and biological heterogeneity possibly over time in order to uncover substructure/subgroups of individuals who have similar underlying trajectories linked to specific profiles of variables representing diverse range of data types covering multiple domains. The statistical models are developed using measurements of cognitive functioning, biomarkers, genetic and environmental data. Additionally, they are used to gain insight into disease processes and to make predictions of an individual's risk of future cognitive decline before symptoms have developed. As early indicated, individual's at high risk are potential candidates for recruitment into clinical trials that test therapies in the very early stages of disease. The models we develop may also provide insight into potential new targets for therapies. These models should be validated before they can be used for recruitment into clinical trials or clinical management of individuals

In the following sections, we describe the datasets requested, obtained and used in WP14B, the specific objectives of WP14B and the research approaches adopted.

- **Datasets**

At the start of the project, we applied through DPUK for access to the following population cohort datasets through the DPUK portal as they had longitudinal cognitive information, had genetics and/or biomarker information and sufficient number of participants for building and validating risk stratification models: The Caerphilly

Prospective Study (CaPS), Cognitive Functioning and Ageing Studies (CFAS I and CFAS II), The English Longitudinal Study of Ageing (ELSA), The European Prospective Investigation of Cancer – Norfolk (EPIC-Norfolk), The Lothian Birth Cohort of 1936 (LBC1936), UK Biobank and Whitehall II Study. By the end of the study, we received the following requested cohorts directly through the portal – CFAS I and CFAS II, ELSA, Whitehall II and Generation Scotland (although not required). CFAS I, CFAS II, ELSA and Generation Scotland were received in March 2018, Whitehall II was received in May 2019 (because of issues with the Data Access Agreement). CFAS I genetic data, obtained through participation in the International Genomics of Alzheimer's Project (IGAP), are not available via the DPUK portal but was externally obtained (alongside CFAS I cohort data) at the beginning of our DPUK involvement. ELSA genetic data linked to the ELSA cohort data had to be applied for through METADAC and this became available within the portal at the start of 2020 after going through the access process for over a year. UK Biobank was externally obtained in August 2018 and arrangement to get its data into WP14B's dedicated DPUK workspace is yet to be completed due to issues regarding size (over 12TB), multiple instances of it (due to for example, updates) and wider DPUK/UK Biobank issues. The Lothian Birth Cohort of 1936, EPIC-Norfolk and CaPs although they can now be directly accessed via the Portal, we have yet to receive within our dedicated DPUK workspace. However, an external request for LBC1936 was made, and access to it became available to us in 2019. In addition to access to DPUK cohort data, we also have access to data from the North American Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (Mueller et al. 2005).

CFAS I and ELSA both have high-dimensional genotyping, cognitive status (e.g. dementia diagnosis), longitudinal cognitive performance and socio-demographic (e.g. age, sex, education) information. However, they do not have imaging data. Whitehall II has longitudinal cognitive performance and socio-demographic information, but no genetic, cognitive status and imaging data. UK Biobank has a wide variety of data modalities at baseline but limited longitudinal information. LBC1936 has longitudinal cognitive performance, cognitive status, socio-demographic, genetic, imaging and metabolomic information. ADNI has all required information but is not a DPUK cohort and the data are collected from the United States of America.

- **Objectives, Analyses and Results**

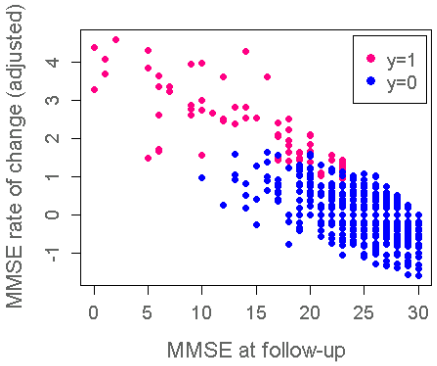
The specific underpinning scientific objectives of WP14B were

a. The need to develop robust (statistical) strategies/algorithms to identify risk stratified subgroups, which can be recruited either into Experimental Studies (including clinical trials) or followed for disease progression based on **cross-sectional information**.
b. The need to develop robust statistical strategies/algorithms for trajectory risk stratification that can identify subgroups of individuals in pre-clinical/prodromal stages of disease who have different progression rates to dementia using **longitudinal information**.
c. The need to statistically identify, incorporate and evaluate varying types and complexities of biomarkers that associate with brain A pathology or with onset of dementias or predictive of treatment response.
d. To provide model-based statistical input, driven by an iterative strategy, which can inform clinical trial designs for dementias-related experimental studies.

To accomplish these goals, we had three main components of our work. The first was based on risk stratification/prediction for rapid cognitive decline: a cross-sectional analysis. The second was trajectory stratification using longitudinal cognitive functioning, imaging, genetic and traditional risk factors to identify subpopulation with differing longitudinal cognitive profiles linked to particular profiles of imaging, genetic and traditional risk factors. The third component was to investigate the potential for designing dementia trials embedded within a cohort.

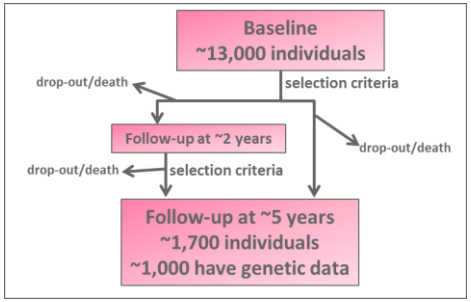
1. ***Risk stratification/prediction for rapid cognitive decline: a cross-sectional analysis***

The aim was to identify subgroups of individuals at high risk of rapid cognitive decline in the short-to-medium term, with a focus on genetic data. We develop the methodology initially using CFAS I, as we had early external access to this cohort, in particular the CFAS IGAP genotyping data. The information used in our analysis was cognitive functioning as measured by the Mini Mental State Examination (MMSE; Folstein, 1975), the Geriatric Mental State Examination's automated geriatric examination for computer assisted taxonomy (AGECAT) for producing a level of confidence in an organicity/dementia "diagnosis" through an algorithmic diagnostic system, baseline risk factors of age, gender, education, marital status and Townsend deprivation index, and the approximately 19,000 "AD-related" SNPs, informed by IGAP (although after quality control filtering the number of SNPs reduced to approximately 9,700). The outcome of interest was a binary variable for rapid cognitive decline over a period of five years after entry into the study derived from MMSE. The rapid cognitive decline category contained the top 15% of individuals with the highest rate of change (adjusted for age, gender and education) and with an MMSE score < 24 at follow-up. The Figure below shows in red those in this category.



Sampling bias

CFAS I had approximately 13,000 individuals recruited at baseline, of whom only a fraction of approximately 1,000 individuals had five-year follow-up and CFAS IGAP genetic data. However, these 1,000 individuals are not representative of the baseline sample due to selection criteria and dropout death. As we are interested in making inference on a representative sample of the baseline population, we calculated sampling weights via inverse probability weighting to account for the fact that our follow-up sample suffers from sampling bias. Logistic regression was used to model the sampling process (see Figure below) with covariates age, gender, education, Townsend deprivation, marital status, centre ID, M<SE and AGECAT included.



Variable selection

To reduce the dimensionality of the CFAS IGAP SNP data we performed variable selection. This was done using logistic regression with either a lasso or elastic net penalty, where we regressed the SNPs on the cognitive decline outcome adjusting for age, gender and education and incorporating the sampling weights into the regression to address the sampling bias issue.

Stratification and Replication

To identify subpopulations (i.e. clusters) of individuals within which individuals share similar cognitive decline and similar genetic profiles on the SNPs selected by our variable selection procedure, we applied a semi-supervised Bayesian mixture (clustering) approach called profile regression (Molitor et al. 2008) using the R package PReMiuM (Liverani et al. 2015). The clusters/sub-populations were identified automatically under this approach. We attempted to validate the findings from profile regression on the CFAS I follow-up data using the independent ADNI data. Unfortunately, we got disagreement between the profiles obtained using the CFAS I data and the profiles obtained using the ADNI data. SNP alleles that seemed to be associated with higher risk clusters in CFAS I were associated with lower risk clusters in ADNI and vice-versa. To further investigate this discrepancy, we carried out a simpler univariate analysis to check assumptions of SNPs with outcome (correcting for age, gender, education and baseline MMSE). Again, we found that there were disagreements between the CFAS analysis results and the ADNI analysis results in the directions of the association between many of the SNPs with small p-values and the outcome. See Figure below. We further checked to ensure that this was not due to different coding of SNPs in the two studies (it was not). In general, there seemed to be a lack of signal using the CFAS IGAP genetic data as exemplified when we compare the association between APOE and the outcome. Here the association using CFAS I was far weaker than the association obtained using ADNI.

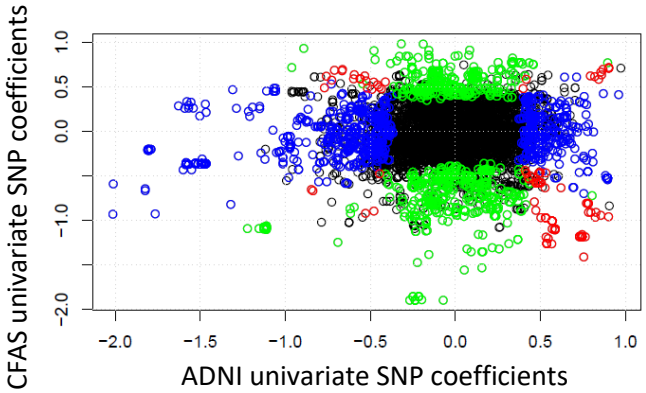


Figure: Regression coefficients from univariate analysis for each SNP using the ADNI data (x-axis) and CFAS data (y-axis). Green points indicate SNPs with p-value less than 0.1 for CFAS but not ADNI. Blue points indicate SNPs with p-value less than 0.1 for ADNI but not CFAS. Red points indicate SNPs with p-value less than 0.1 in both datasets.

Polygenic risk score (PRS) on CFAS I data

Instead of aiming to stratify individuals based on a set of SNPs obtained through variable selection, we instead use information from across all the SNPs by encoding into a polygenic risk score. To construct a polygenic risk score we explored two approaches using 777 individuals in CFAS I with data on 9035 SNPs. The first approach was based on estimating the individual SNP weights for the PRS from the CFAS I data using the polygenic risk score methodology adopted by Escott-Price et al. (2015) and previously described by the International Schizophrenia Consortium (Purcell et al. 2009). The second approach used the weights derived from the IGAP study.

Approach 1 proceeded as outlined below

- 5-fold cross-validation (CV) approach
- To estimate the SNP weights, logistic regression was used on training data adjusting for first 10 principal components of the SNPs, baseline MMSE, age, gender, education
- PRS then calculated for individuals in test data, using SNPs with p-value <= 0.05, and logistic models fitted with covariates:
  Model 1: PRS only
  Model 2: PRS + APOE4
  Model 3: PRS + APOE4 + age + gender
  Model 4: APOE4 + age + gender
- Predictions also evaluated using the test data.
- Area under the ROC curves, AUCs (mean +/- SD across the 5 train/test splits):
  Model 1: 0.60 +/- 0.10
  Model 2: 0.58 +/- 0.12

Model 3: 0.70 +/- 0.08
Model 4: 0.67 +/- 0.08

We found weak evidence that the PRS derived under Approach 1 improved prediction when added to APOE4 genotype, age and gender (AUC of 0.70 vs 0.67).

Approach 2 proceeded as outlined below

- PRS calculated for individuals in training data and logistic models fitted as above.
- Predictions evaluated using the test data
- AUCs (mean +/- SD across the 5 train/test splits):
  Model 1: 0.56 +/- 0.05
  Model 2: 0.54 +/- 0.05
  Model 3: 0.67 +/- 0.08
  Model 4: 0.68 +/- 0.09

We found that the PRS derived using the IGAP weights had weak predictive power, but no evidence that it improves over a model with just the number of APOE4 alleles, age and gender (AUC of 0.67 vs 0.68).

In both Approach 1 and Approach 2, Model 2 did not outperform Model 1. Thus the inclusion of number of APOE4 alleles in addition to the PRS does not improve predictive performance even though APOE4 is not contained in the PRS.

As an additional investigation, we examined the extremes of the PRS distribution under both approaches. For the PRS with weights obtained using the CFAS I data, among individuals in the top 20% of the PRS distribution, 9% +/- 5% are cases (i.e. have rapid decline according to our outcome variable). Here the mean and SD are over the CV test sets. Amongst individuals in the bottom 20% of the PRS distribution, 7% +/- 4% are cases. Therefore there was no clear difference seen in the proportion of cases between the upper quintile and the lower quintile of the PRS distribution.

For the PRS with IGAP weight, 10% of the individuals in the upper quintile of the PRS distribution are cases, whilst 4% are cases in the bottom quintile of the PRS distribution. These proportions are calculated using all 777 individuals. The overall proportion that are cases is 8%. Here there is some indication of a difference with the highest quintile being somewhat enriched for cases.

Polygenic risk score (PRS) on ADNI data

When similar strategy (as described earlier) was employed on 515 individuals recruited into ADNI with data on 8515 SNPs, we got the below results for Approach 1 (estimating PRS weights from ADNI data) and Approach 2 (IGAP weights).

Approach 1 (estimating PRS weights from ADNI data):

- 5-fold cross-validation (CV) approach

- To estimate the SNP weights, logistic regression was used on training data adjusting for first 10 principal components of the SNPs, baseline MMSE, age, gender, education
- PRS then calculated for individuals in test data, using all SNPs and logistic models fitted with covariates:
  Model 1: PRS only
  Model 2: PRS + APOE4
  Model 3: PRS + APOE4 + age + gender
  Model 4: APOE4 + age + gender
  Model 5: APOE4 + age + gender + imaging
  Model 6: PRS + APOE4 + age + gender + imaging
- Imaging variables used: baseline ventricles and whole brain volumes, both normalised by intracranial volume (ICV)
- Predictions also evaluated using the test data.
- AUCs (mean +/- SD across the 5 train/test splits):
  Model 1: 0.56 +/- 0.08
  Model 2: 0.69 +/- 0.04
  Model 3: 0.74 +/- 0.05
  Model 4: 0.71 +/- 0.06
  Model 5: 0.78 +/- 0.09
  Model 6: 0.81 +/- 0.07

We found PRS alone has some predictive power and that there is some evidence (a bit stronger than for CFAS I) that it improves prediction when added to number of APOE4 alleles, age and gender (Model 3 vs Model 4) and also when added to number of APOE4 alleles, age, gender and imaging (Model 6 vs Model 5). It was observed that, unlike when using CFAS data, that the addition of APOE4 increased the predictive performance (Model 2 vs Model 1). The addition of the baseline imaging brain volume variables of ventricles and whole brain (normalised by ICV) also showed clear improvement in predictive performance (Model 5 vs Model 4, and Model 6 vs Model 3).

Approach 2 (IGAP weights):

- PRS calculated for individuals in training data and logistic models fitted as above.
- Predictions evaluated using the test data
- AUCs (mean +/- SD across the 5 train/test splits):
  Model 1: 0.62 +/- 0.13
  Model 2: 0.68 +/- 0.11
  Model 3: 0.66 +/- 0.08
  Model 4: 0.64 +/- 0.08
  Model 5: 0.74 +/- 0.11
  Model 6: 0.73 +/- 0.10

Similar findings as for Approach 1 are obtained, although the benefits of PRS over other variables are less clear. A caveat here is that the IGAP weights were estimated from a meta-analysis of many cohorts, including ADNI.

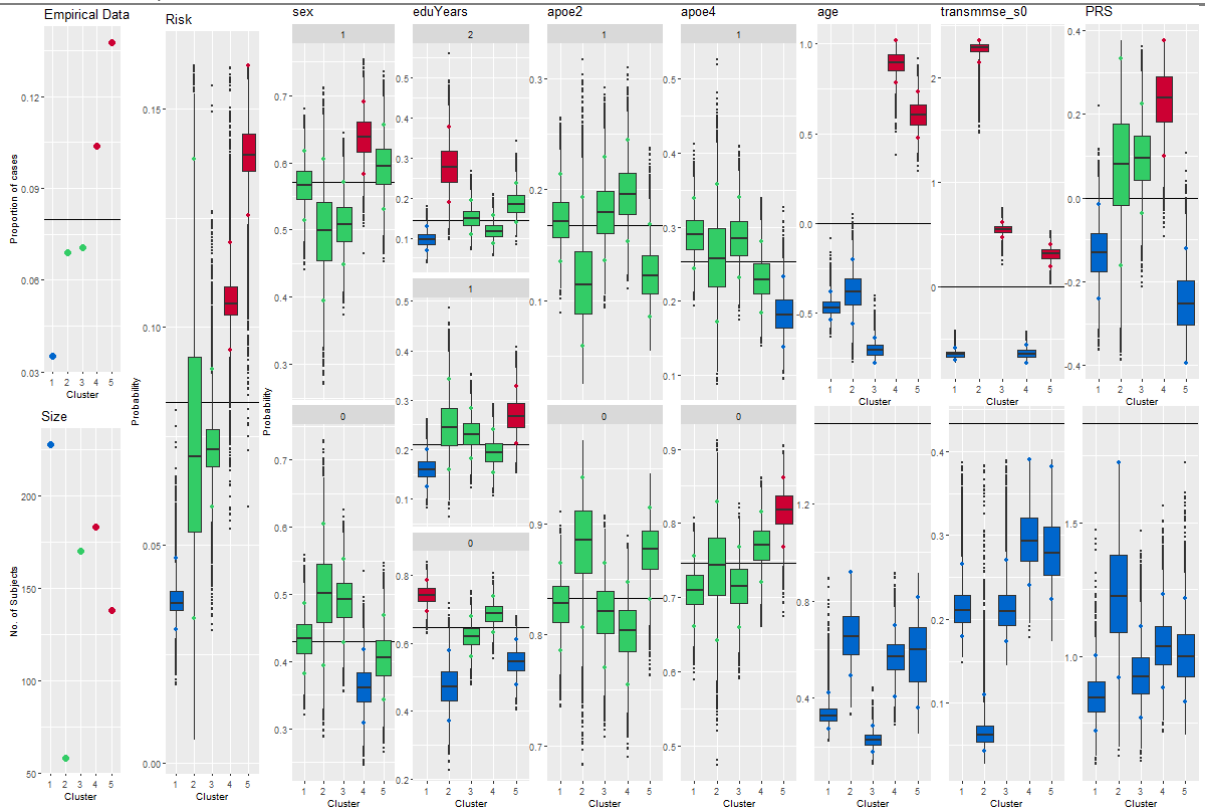
A look at the extremes of the PRS distribution under Approach 1 found that among individual in the top 20% of the PRS distribution, 11% +/- 8% are cases having rapid cognitive decline). Here, the mean and SD are over the CV test sets. Among the bottom quintile of the PRS distribution, 8% +/- 3% are cases. Again, there was no clear evidence of enrichment from being in the top quintile of the PRS distribution. Under Approach 2, 17% of individuals in the top quintile were cases, while 4% were cases in the bottom quintile. Here there was clear evidence of enrichment.

When the PRS obtained using ADNI data was used for prediction using CFAS I data, the area under the ROC curve was 0.52. Further looking at the extremes of this PRS distribution in CFAS I provided no evidence for enrichment of cases in the top quintile versus the bottom quintile (9% vs 7%).

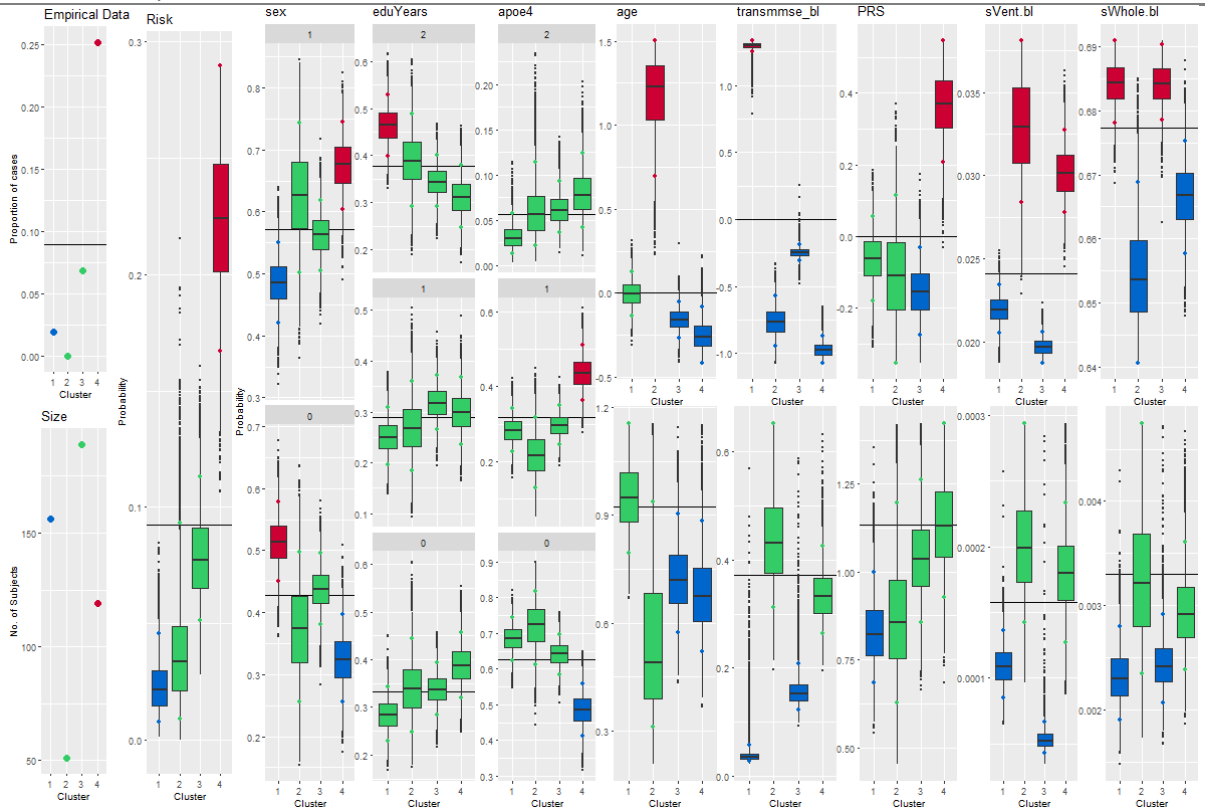
Profile regression using CFAS I and ADNI with IGAP PRS instead of individually selected SNPs

On applying profile regression to the binary rapid decline cognitive outcome and the variables age, gender, education, baseline MMSE, APOE4 positivity, APOE2 positivity and the IGAP PRS using CFAS I data, we obtained the 5 clusters described in the Figure below. Clusters 4 and 5 represent higher-risk subpopulations. Cluster 4 is characterised by higher PRS scores, while cluster 5 has lower PRS scores and less APOE2 and APOE4 positive patients. Both clusters are enriched for older individuals.

WP 14b End report



The results of Profile regression applied to the ADNI data of rapid cognitive decline, age, gender, education, baseline MMSE, APOE4 genotype ,IGAP PRS and the brain imaging volumes of ventricles and whole brain volume is displayed in the Figure below.



We found 4 clusters. One of these clusters, Cluster 4, was a higher-risk subpopulation enriched for higher PRS scores, APOE4 positivity, males, lower baseline MMSE, higher ventricle volume and lower whole brain volume. Cluster 1, the lowest risk cluster has the opposite characteristics to Cluster 5. Additionally, there is another lower risk cluster (Cluster 2) that is enriched for higher ventricle volume and lower whole brain volume. This cluster may be indicative of a group of individuals with AD pathology yet still high cognitive functioning, giving some support to the cognitive reserve hypothesis of Stern (2009) that some individuals may have more ability, using compensatory processes, to cope with brain damage, delaying decline in cognitive functioning and onset of cognitive symptoms. A further illustration of this hypothesis is seen in the longitudinal trajectory stratification analysis of the next section.

Future work on risk stratification will be to validate the clustering obtained from applying profile regression with PRS in ADNI/CFAS I to either or both of ELSA and LBC1936.

2. *Trajectory stratification/prediction for cognitive decline: a longitudinal analysis*

WP 14b End report

The aim here was to develop and apply novel methodology for trajectory stratification of heterogeneous clinical and biological data that integrate longitudinal outcome with biomarker information and risk factors in order to identify subpopulations/clusters of individuals. The methodology developed is an extension of the profile regression methodology of Molitor et al. (2010) mentioned above, but now further incorporating a longitudinal outcome instead of a univariate outcome. This extension is described fully in Chapter 11 (Rouanet, Richardson and Tom, 2020) of the book "Bayesian Methods in Pharmaceutical Research", edited by Lesaffre, Baio and Boulanger (see https://www.crcpress.com/Bayesian-Methods-in-Pharmaceutical-Research/Lesaffre-Baio-Boulanger/p/book/9781138748484).

Chapter11-proofs.pdf

Briefly, our work focuses on the study of cognitive decline and brain imaging for precision/stratified medicine and uses data on a subset of 199 patients from the ADNI cohort as an exemplar dataset. As longitudinal cognitive, neuroimaging, genetic and socio-demographic information were not initially available to us through the DPUK portal, we plan in future to apply our novel statistical methodology to the Lothian Birth Cohort of 1936.

Note that, in general, the identification of clusters of subjects with different susceptibility to cognitive decline based on repeated cognitive measurements and baseline neuroimaging volumetric information, APOE4 carrier status, gender and education from observational data, such as ADNI, is challenging. Firstly, there is no clinical evidence regarding the number of clusters to be expected. Additionally, the *a priori* specification of particular types of longitudinal cognitive trajectories can strongly influence the clusters obtained. Finally, interpretation of results should also account for the uncertainty in any clustering structure arrived at from analysing the data, as the true underlying subpopulations are unknown. The profile regression extension for longitudinal outcome we describe in Chapter 11 of the aforementioned book is done within a flexible nonparametric Bayesian framework, where Dirichlet Process priors are used to deal with the unknown number of cluster and the propagation and quantification of the uncertainty of clustering allocation, and Gaussian Process priors are used to flexibly model the longitudinal outcome so as to not be restricted to particular *a priori* parametric forms of cognitive trajectories.

The results of applying the methodology to the exemplar dataset of 199 individuals from the ADNI cohort is shown in the two Figures below. The longitudinal outcome was MMSE normalised using the monotonic transformation of Philipps et a. (2014). The baseline neuroimaging markers considered in the analysis measured the volumes of six cerebral regions using MRI. The six regions were the ventricles, which include four communicating cavities producing the cerebrospinal fluid; the hippocampus, a ridge of grey matter tissue involved in neurogenesis in adults; the entorhinal cortex, an outer layer of grey matter; the fusiform and the mid-temporal gyri, which are folds in the cortex; and the whole brain. These regions are known to play key roles in memory (Kempermann et al. 2015; Breteler et al. 1994).

We identify four clusters through our extended profile regression analysis, representing four subpopulations with differing profiles of normalised MMSE and brain volumes.

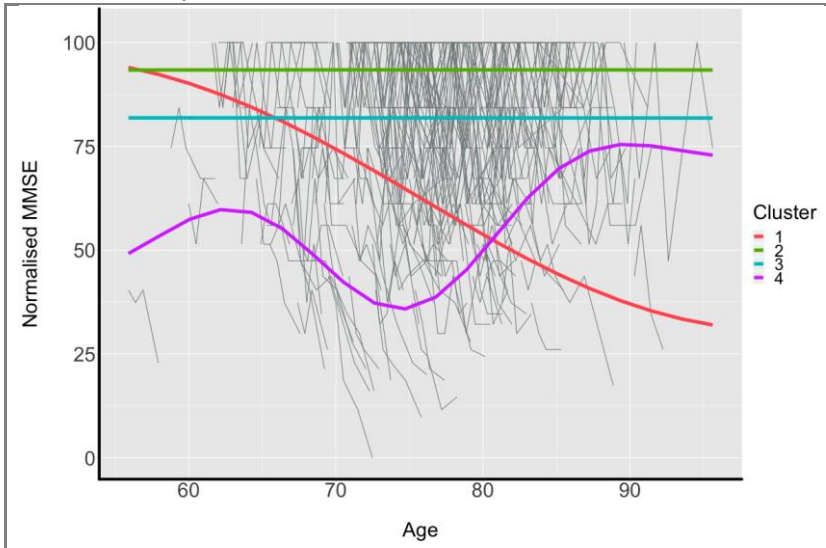
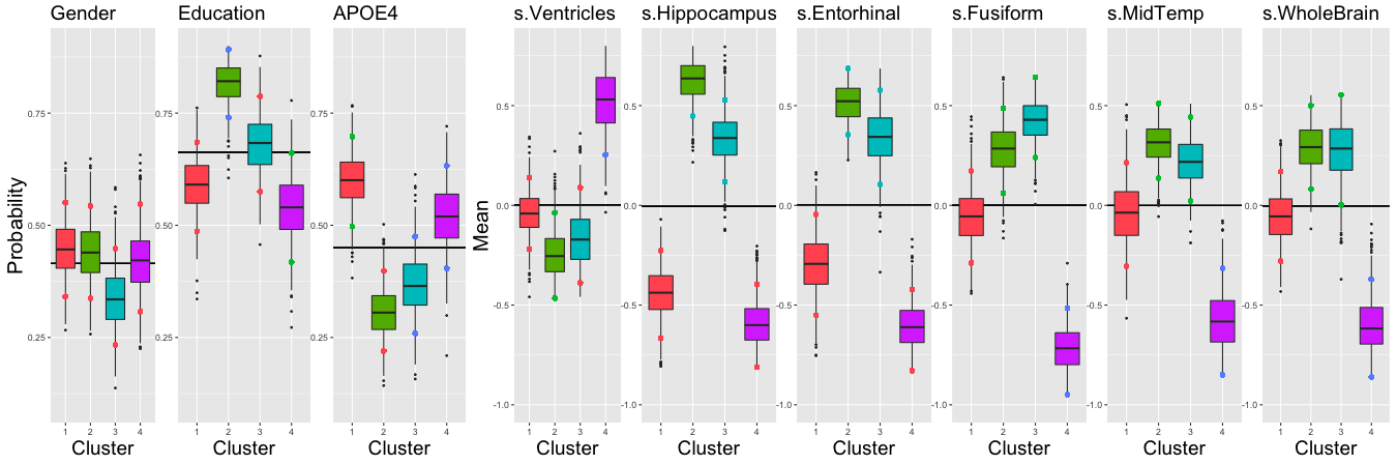


Figure: The four class-specific normalised MMSE cognitive trajectories obtained from the extended profile regression analysis (coloured lines), with the observed trajectories (grey lines) in the sample of 199 individuals from ADNI

Figure: The socio-demographic, genetic and volumetric imaging profiles the four representative clusters from the extended profile regression analysis



Cluster 1 (in red) is characterised by subjects who on average have steep cognitive decline, “average” volumes of ventricles, fusiform, mid-temporal gyri and whole brain and low levels of hippocampal and entorhinal cortex volumes. Moreover, in this cluster, there is a high proportion of APOE4 carriers and a relatively low percentage of individual with high levels of education.

Clusters 2 and 3 represent cognitively unimpaired individuals, with stable cognitive trajectories differing only by baseline levels of cognitive functioning. The imaging profiles of these two clusters are similar, with, on average, low ventricle volumes, and high volumes for hippocampus, entorhinal cortex, fusiform, mid-temporal gyri and whole brain. Cluster 2 is enriched for individuals with higher levels of education and non APOE4 carriers.

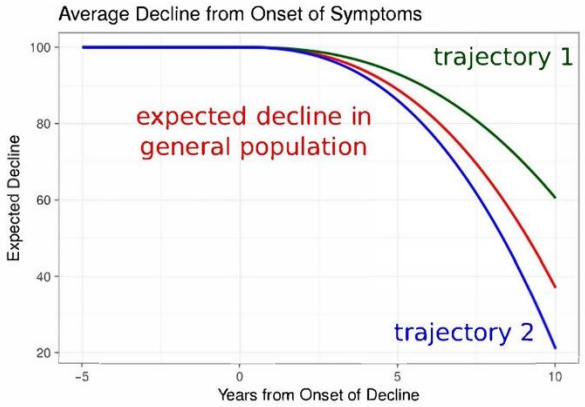
Cluster 4 displays a peculiar cognitive trajectory pattern. Further investigations revealed that this cluster actually comprises of two subgroups: a group with steep cognitive decline who were followed-up between ages 65 and 80 years old, and a group with a more stable cognitive profile who was observed from 80 years old onwards. These two subgroups, however, have similar imaging, genetic and socio-demographic profiles; illustrating again the possibility for the cognitive reserve hypothesis of Stern (2009) to be at play.

Cluster 1 and the first subgroup of Cluster 4 represent “high-risk” subpopulations of individuals for whom recruitment into AD clinical trials should be targeted.

3. Designing dementia trials embedded within a cohort

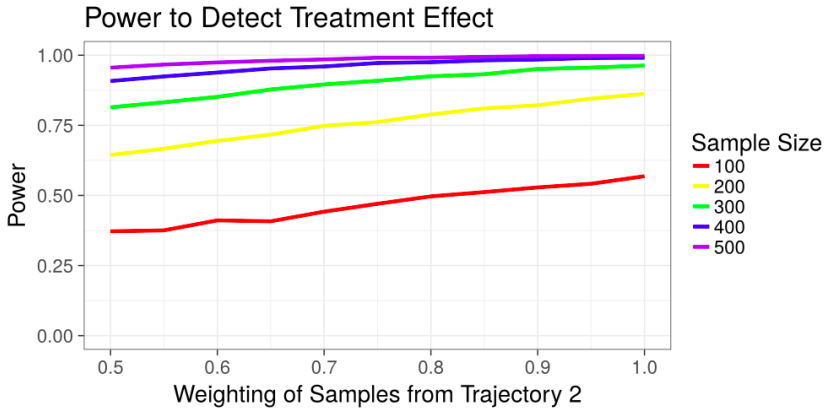
The running of trials within cohorts enables the use of information already collected on individuals to be used to inform selective recruitment and adaptation of clinical trials. It can also increase efficiency and aid indirect treatment comparison. We investigate how to sample from a single disease cohort to form a trial population to balance the desire to estimate a trial treatment effect with say 80% power and conditional on this to estimate an extrapolated treatment effect in the cohort population with the highest precision.

We inform the design of our simulation study from the work done within EPAD. We use the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) Total Index Score (Randolph et al. 1998) as the trial’s primary outcome variable and we assume that the effect of treatment is multiplicative on the rate of decline. We assume that the disease cohort is made of a mixture of subpopulations with differing disease trajectories and from which we wish to recruit those whose trajectory would enable the detection of a treatment effect. However, the subpopulation in which an individual belongs is unknown, we can only estimate the probability that an individual belongs to a certain subpopulation based on information accumulated on the individual whilst in the cohort. For our simulation work, we assume that the longitudinal trajectory data for individuals in the cohort is generated from a latent class mixed model (LCMM) with a fixed number of classes (and to begin with we assume two) that represent the different underlying disease trajectory subpopulations. The Figure below provides an example of the possible underlying trajectories in the two subpopulations (trajectory 1 in green and trajectory 2 in blue is the faster decliner subpopulation) and overall in the full cohort (in red).

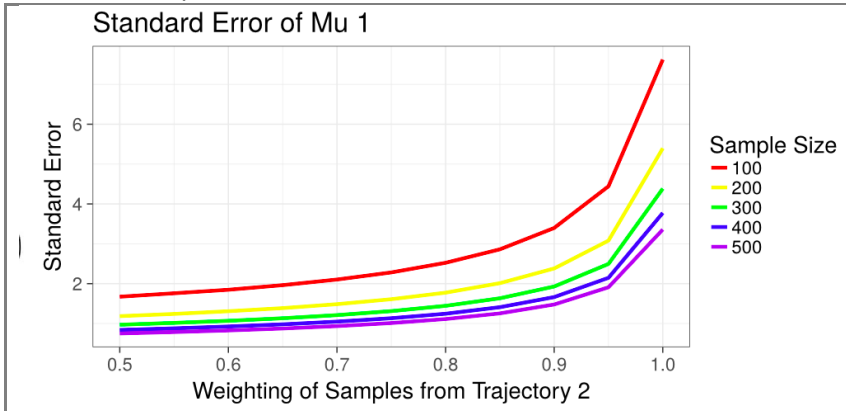


From the individuals in our simulated cohort, we fit a latent class mixed model using the R package lcmm (Proust-Lima at al. 2017) to their longitudinal data. We estimate the posterior probabilities of belonging to the two subpopulations from this latent class mixed model and use these probabilities for each individual, with appropriate weighting, to derive an individual's sampling probability for inclusion into the trial. If sampled, the individual is randomised to receive either treatment or control in a particular ratio (typically 1:1) in the standard way and the appropriate observed RBANS outcome over the follow-up period for this individual in the clinical trial is accordingly generated depending on trial arm and design specifications of the trial.

The Figure below shows the result of a particular simulation scenario to determine the power to detect a treatment effect over different sample sizes. The sample sizes are for illustrative purposes and should not be indicative of the actual sizes of clinical trials in AD. The patterns are more relevant.



From this Figure, one observes that as the we increase the weight of favouring individuals from Trajectory 2 (faster decliner subpopulation) over Trajectory 1 (the slower decliner subpopulation) there is an increased power of detecting a treatment effect if it truly exists. However, this comes at the cost of decrease precision in estimating the subpopulations' rate of decline for the cohort as shown in the below Figure, where "Mu 1" represents the rate of decline in the subpopulation with Trajectory 1.



Thus, we need to determine a strategy which prioritises reaching some power threshold, and uses and any additional samples to minimise the standard error of the subpopulation estimates for the cohort. This work can be extended to more than two subpopulations, to different measures of cohort-level effects and to multiple correlated outcomes instead of a single composite outcome such as RBANS Total.

The attached document has further details of this work.

Trial Embedded within Cohort - Fort

- **Conclusion**

In conclusion, we have investigated statistical methodologies for risk and trajectory stratification using either cross-sectional or longitudinal outcome data and data from other modalities such as genetics, neuroimaging and socio-demographic risk factors. The methodology focuses on identifying subpopulations of individuals with varying risk (in particular those of high risk) of rapid cognitive decline (as a surrogate of higher risk of onset of dementia). However, instead of capturing risk through a single summary measure, we instead jointly model the various pieces of information related to risk (e.g. cognitive functioning, genetics, MRI volumes, age, sex, education) to better characterise the complex multi-dimensional risk spectrum. By adopting such methodology, we hope to gain more insight into the disease process and increase efficiency of clinical trials by further tailoring recruitment. Moreover, this methodology can be used for prediction and therefore help aid clinical management decisions. Although we have not yet investigated using this methodology in designing dementia trials embedded within a cohort, we have investigated using latent class mixed modelling methodology for this purpose, in particular when there are competing roles of powering a trial and estimating cohort-specific effects.

References

- Breteler MM, van Amerongen NM, van Swieten JC, Claus JJ, Grobbee DE, Van Gijn J, Hofman A, van Harskamp F. Cognitive correlates of ventricular enlargement and cerebral white matter lesions on magnetic resonance imaging. The Rotterdam Study. Stroke. 1994; 25(6):1109-15
- Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, Badarinarayan N, GERAD/PERADES, IGAP consortia, Morgan K, Passmore P. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain. 2015; 138(12):3673-84.
- Folstein M, Folsten S, McHugh P.Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. J Psychiatry Res. 1975; 12:189–198.
- Kempermann G, Song H, Gage FH. Neurogenesis in the adult hippocampus. Cold Spring Harbor perspectives in biology. 2015; 7(9):a018812.
- Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM: An R package for profile regression mixture models using Dirichlet processes. Journal of statistical software. 2015; 64(7):1.
- Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics. 2010; 11(3):484-98.
- Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia. 2005; 1(1):55-66.
- Philipps V, Amieva H, Andrieu S, Dufouil C, Berr C, Dartigues JF, Jacqmin-Gadda H, Proust-Lima C. Normalized mini-mental state examination for assessing cognitive change in population-based brain aging studies. Neuroepidemiology. 2014; 43(1):15-25.
- Proust-Lima C, Philipps V, Liquet B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. Journal of Statistical Software. 2017; 78(i02).
- Purcell SM, Wray NR, Stone JL, Visscher PM. O0Donovan MC, Sullivan PF et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748-52.
- Randolph C, Tierney MC, Mohr E, Chase TN. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity. Journal of clinical and experimental neuropsychology. 1998; 20(3):310-9
- Stern Y. Cognitive reserve. Neuropsychologia. 2009 Aug 1;47(10):2015-28.
- World Health Organization. Towards a dementia plan: a WHO guide. 2018