WP 14B

Biostatistics – methods
Start date: 1 March 2016. Completion date: 30 March 2020

Overall work package 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.

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 in to models	M1.1.3 Dec 2019		
	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

Updates on delivery against milestones since last report

- **M1.1.3 Integration of imaging data in to models**

Analysis using the Lothian Birth Cohort is still underway. The analysis with the ADNI dataset is completed. The delay is due to Anais Rouanet leaving the Unit at the end of August 2019 to take up a new research position in Bordeaux. She has agreed to finish off this work whilst in Bordeaux.

- **M1.1.4 Paper submitted to peer review journal**

Paper still to be finished as the analysis using the Lothian Birth Cohort data is yet to be completed.

- **M3.1.1 Strategy in place and statistical activity underway**

Strategy is in place but analysis is still underway due to the departure of Anais Rouanet.
Whitehall II data are accessible to us from the DPUK portal. Our preliminary investigations of the SNP data overlap with the CFAS SNP data obtained through I-GAP showed that there were limited overlap and therefore Whitehall II was not an ideal candidate to validate our CFAS I cohort analysis results where we explored whether subgroups at high (or low) risk of rapid cognitive decline can be identified using genetic data.

Summary of plan to deliver on outstanding work (with dates)

Investigation will continue into the value of incorporating both polygenic risk scores and imaging biomarkers into risk stratification and trajectory risk stratification work on identifying subgroups with different cognitive trajectories of transition rates to dementia.
Applications of the various methodologies to data from accessible DPUK cohorts, including linked ELSA clinical and genetic data
Write up of the various pieces of work when analyses completed.

Team members funded (full or part-time) by DPUK
Anais Rouanet and Steven Hill (**no longer employed on DPUK funding**)
Team members involved with the project but not funded by DPUK
Brian Tom, Sylvia Richardson

Risks	**Mitigation**
1) N/A	1)

Outcomes

1. Publications for WP14B

Rouanet, A, Richardson, SR, and Tom, BD (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

chapter 11.pdf

Early proof (confidential):

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

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.

2. Additional Outputs

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 are currently using the Lothian Birth Cohort data.) We have presented the work at various meetings including at DPUK and ARUK conferences.

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.

Project narrative

Since becoming a DPUK partner on the 1st March 2016, the MRC Biostatistics Unit has been developing statistical methodology for baseline risk stratification and trajectory stratification using cognitive, genetic and imaging data. Using the CFAS I cohort, we have used Bayesian profile regression to explore whether subgroups at high (or low) risk of rapid cognitive decline can be identified using SNP data. This is particularly difficult task as SNP-specific genetic effect sizes in dementia research are generally small and therefore difficult to detect and so has led to the use of polygenic risk scores over the use of individual SNPs. We have attempted to validate our findings using Whitehall II Study. Unfortunately, the overlap of SNP data from Whitehall II and CFAS I is not as large as originally hoped. Thus Whitehall II cannot be used to validate the findings obtained from CFAS I. We are exploring whether the ELSA cohort can be used for validation. Nevertheless, additional analyses, beyond validation, are planned using Whitehall II and ELSA data.

We have written a book chapter that looks at the benefit of Bayesian outcome-driven clustering for longitudinal data in precision medicine, with application to Alzheimer's disease. Here we have investigated both latent class mixed models and profile regression extended to handle longitudinal MMSE data to identify subpopulations with latent cognitive trajectories that are linked to specific MRI volumetric imaging (e.g. ventricles, hippocampus, entorhinal cortex, fusiform gyrus, middle temporal gyrus and whole brain) and risk factor (e.g. APOE4 status, education and gender) profiles. This work may allow early identification, based on both covariate and outcome data, of subjects who are expected to have stable disease course and different rates of progression of disease and the future possibility of either intervening early on in disease to alter disease course or informing early management strategy. It will also provide a reference for researchers to state-of-the-art statistical methodology that can be used for precision medicine in dementia research.

In addition to the book chapter, we have developed and made available to researchers R code that allows Bayesian profile regression to be used for longitudinal data. Also we are exploring how this profile regression method can be extended to incorporate biomarkers from other types of modalities and biomarkers changing over time.