Quantifying cognitive and mortality outcomes in older patients following acute illness using epidemiological and machine learning approaches

Candidate: Dr Alex Tsui

PhD Thesis 2022, Institute of Cardiovascular Sciences, University College London (UCL)
Supervisors: Prof Daniel Davis, Prof Nishi Chaturvedi, Prof Parashkev Nachev

Declaration

I, Alex Chun Kong Tsui, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Introduction

Cognitive and functional decompensation during acute illness in older people are poorly understood. It remains unclear how delirium, an acute confusional state reflective of cognitive decompensation, is contextualised by baseline premorbid cognition and relates to long-term adverse outcomes. High-dimensional machine learning offers a novel, feasible and enticing approach for stratifying acute illness in older people, improving treatment consistency while optimising future research design.

Methods

Longitudinal associations were analysed from the Delirium and Population Health Informatics Cohort (DELPHIC) study, a prospective cohort ≥70 years resident in Camden, with cognitive and functional ascertainment at baseline and 2-year follow-up, and daily assessments during incident hospitalisation. Second, using routine clinical data from UCLH, I constructed an extreme gradient-boosted trees predicting 600-day mortality for unselected acute admissions of oldest-old patients with mechanistic inferences. Third, hierarchical agglomerative clustering was performed to demonstrate structure within DELPHIC participants, with predictive implications for survival and length of stay.

Results:

i. Delirium is associated with increased rates of cognitive decline and mortality risk, in a dose-dependent manner, with an interaction between baseline cognition and delirium exposure. Those with highest delirium exposure but also best premorbid cognition have the “most to lose”.
ii. High-dimensional multimodal machine learning models can predict mortality in oldest-old populations with 0.874 accuracy. The anterior cingulate and angular gyri, and extracranial soft tissue, are the highest contributory intracranial and extracranial features respectively.

iii. Clinically useful acute illness subtypes in older people can be described using longitudinal clinical, functional, and biochemical features.

Conclusions

Interactions between baseline cognition and delirium exposure during acute illness in older patients result in divergent long-term adverse outcomes. Supervised machine learning can robustly predict mortality in in oldest-old patients, producing a valuable prognostication tool using routinely collected data, ready for clinical deployment. Preliminary findings suggest possible discernible subtypes within acute illness in older people.

Impact Statement

Acute illness in older people commonly results in decompensation of premorbid function, cognition and increased mortality risk. Delirium, an acute confusional state resulting from cognitive decompensation, is distressing to patients and carers, associated with long-term adverse consequences. Individualised outcome prediction using low-dimensional models in this population group had been inaccurate and unreliable, with poor reproducibility between geographical and healthcare settings. Stratification by clinical presentations with divergent consequential recovery trajectories had not been robustly articulated, with acute illness of older people commonly recognised as a single entity in clinical practice.

Academically, these findings first advance the definition of delirium to reflect a longitudinal construct, which can only be fully understood by accounting for non-linear interactions with premorbid cognition. In addition, delirium exposure should be quantified as a cumulative dose instead consideration as incidence alone. Second, these findings suggest delirium may be the most appropriate currently available marker of acute illness severity in older people: it is unclear why delirium causes the most deleterious cognitive sequalae in those with the best baseline cognition, but it is most plausible that quantification using physiological measures alone is suboptimal in this patient group. Third, the high prevalence of delirium on discharge emphasises the importance to capture delirium outside hospital settings in future research studies, into the community, if the true burden of delirium is to be accurately quantified.

The high fidelity of the long-term mortality model, as well as identification of structure in describing acute illness within PISCA, are convincing demonstrations of concept for applying machine learning prediction and inferential models in older patients. By highlighting extracranial soft tissue, angular gyri and anterior cingulate as highly contributory anatomical areas, future potential interventions may be targeted towards optimisation of sarcopenia, motor function and hypothalamic-pituitary-adrenal axis respectively.

This thesis showed that discernible patterns of decompensation, demonstrable at a population and individualised levels, exist in older people during acute illness, with clinically relevant consequences. Deeper phenotyping of acute illness in older patients offer wide-reaching clinical potentials: treatments can be guided towards a curative or palliative intent, discharge planning can be aided by accurate prediction of likely new cognitive baseline following acute illness, while it will be possible to identify patients who will benefit most from cognitive follow-up and rehabilitation post-delirium. For future clinical trials of candidate prevention or interventions strategies, the most appropriate patient cohorts can be selected to maximise demonstration of efficacy.

Clinical deployment of stratification strategies can vary in the level of phenotypic granularity employed: at the broadest level, findings from the population epidemiology elements of this thesis utilise baseline cognition and cross-sectional delirium markers. With the most high-dimensional, multi-modal approach, inclusion of investigations such as neuroimaging can produce highly accurate individualised prognostication. Middle-ground strategies with lower-dimensional models or pre-determined clusters, using readily available bloods and clinical features, may offer a pragmatic compromise for immediate deployment across the health service. I demonstrate the potential of investing in investigation modalities and computing performance: accurate stratification can lead a paradigm shift in better targeting of community, pre-admission triage and inpatient treatments.

Acknowledgements

My first thanks must be to Daniel Davis: supervisor, mentor, colleague and friend, whose delirium, geriatric and statistical knowledge knows no bounds, who has developed me as a clinician, epidemiologist and person immeasurably. Never has a chance meeting at the Baker Street Everyman cinema been so academically fruitful.

My thanks also to Parashkev Nachev, constantly inconceivably generous with his polymathic wisdom, patiently improving me from an amateur data scientist to a slightly less amateur data scientist.

Petru-Daniel Tudosiu, my coding and data science brother, with his unfailingly kind nature, has over the course of our years of collaboration provided more solutions to coding crises than either of us could count.

Of course, my heartfelt appreciation to the participants of the DELPHIC study, who generously volunteered their time to make such rich science possible, as well as my DELPHIC colleagues who worked on collecting and curating this incredible study. My colleagues at the MRC Unit for Lifelong Health and Ageing, in particular Nishi Chaturvedi and Marcus Richards, have enriched my knowledge of life-course and population epidemiology over the years.

My biggest thanks are of course for Marissa, the most wonderful long-suffering wife, mother, best friend and delirium expert by osmosis. Any humble achievements are always shared: you are the foundation for all that we build together.

For Marissa, goose and emu

Table of Contents

Declaration 2

Abstract 3

Impact Statement 5

Acknowledgements 7

1 Clinical Case 12

2 Completed publications to date within PhD 14

3 Introduction 16

3.1 Clinical Importance of Delirium 17

3.1.1 Epidemiology of delirium 17
3.1.2 Adverse Outcomes Associated with Delirium 18
3.1.3 Long-term economic costs 19

3.2 Delirium diagnosis 19

3.2.1 Practical issues with delirium screening tools 20
3.2.2 Changing reference standard and validity 21
3.2.3 Operationalising delirium definitions 22
3.2.4 Delirium superimposed on dementia 24

3.3 Management of Delirium 26

3.3.1 Delirium prevention 26
3.3.2 Delirium aetiologies 26
3.3.3 Potential mechanisms underlying delirium 27

3.4 Summary 28

3.5 Aims and objectives 30

3.6 Hypotheses 31

4 Methods 32

4.1 Statistical methods 33

4.1.1 Linear regression and regression splines 33
4.1.2 Kaplan Meier and Cox proportional hazards 34
4.1.3 Machine learning 35
4.1.4 Supervised learning methods 38
4.1.5 Hyperparameter tuning 46
4.1.6 Feature importance 47

4.2 Clustering 47

4.2.1 Gower matrix 47
4.2.2 T-stochastic neighbour embedding (TSNE) 49

4.2.3 Clustering methods 50
4.3 Statistical Parametric Mapping (SPM) 52
4.3.1 What is SPM? 52
4.3.2 Pre-processing 52
4.3.3 Statistical Parametric Mapping 53
4.4 Datasets 56
4.4.1 Delirium and Population Health Informatics Cohort (DELPHIC) study 56
4.4.2 UCLH Cognitive Status Dataset 59
5 Delirium and Population Health Informatics Cohort (DELPHIC) study 64
5.1 Introduction 65
5.2 Methods 67
5.2.1 Ascertainment of delirium 67
5.2.2 Statistical analyses 67
5.3 Results 71
5.3.1 Demographics 71
5.3.2 Delirium prevalence 75
5.3.3 Baseline association with delirium measures 77
5.3.4 Delirium measures and long-term adverse outcomes 92
5.4 Discussion 106
5.4.1 Summary of findings 106
5.4.2 Significance and reasons for findings 106
5.4.3 Alignment with current literature 107
5.4.4 Strengths and limitations 109
5.4.5 Summary and future directions 110
6 Individualised patient prediction of mortality at 600 days post-discharge 112
6.1 Chapter Introduction 113
6.2 Methods 114
6.2.1 Kaplan Meier and Cox Proportional Hazards 115
6.2.2 XG boost classifier model 115
6.2.3 Anatomical inference 116
6.3 Results 116
6.3.1 Study summary 116
6.3.2 Mortality prediction from multimodal data 122
6.3.3 Anatomical correlates of mortality 128
6.3.4 Anatomical features of predictive importance 128
6.4 Discussion 133
6.4.1 The predictability of mortality from routine clinical data 133
6.4.2 Possible mechanisms of increased mortality 135
6.4.3 A multimodal index of frailty 136
6.4.4 Strengths and limitations 137
6.4.5 Conclusion 138

7 Precise Identification of delirium Subtypes through Clinical Analyses (PISCA)..... 140

7.1 Rationale for clustering..... 140

7.2 Methods..... 142

7.2.1 Data pre-processing..... 143
7.2.2 Autocorrelation..... 143
7.2.3 Two dimensional manifold..... 143
7.2.4 Cluster selection..... 143
7.2.5 Model prediction..... 144
7.2.6 Sensitivity Analyses..... 145

7.3 DELPHIC clustering results..... 146

7.3.1 Autocorrelation..... 146
7.3.2 Perplexity Selection..... 148
7.3.3 Cluster size..... 150
7.3.4 Survival prediction..... 153
7.3.5 Length of stay prediction..... 157

7.4 Conclusions..... 159

7.4.1 Presence of structure during acute decompensation of older patients..... 159
7.4.2 Importance of clustering..... 160
7.4.3 Strengths and limitations..... 160

8 Conclusions..... 163

8.1 What are the strengths of this thesis?..... 163

8.1.2 Datasets..... 166

8.2 Ongoing limitations and challenges across the research field..... 167

8.2.1 Study design and sampling..... 167
8.2.2 Outcome ascertainment..... 169
8.2.3 Lack of multimodality, granularity and prosepective capture in covariate ascertainment 170
8.2.4 Questions for future research..... 171
8.2.5 Novel approaches to future studies..... 174

8.3 The Future..... 176

UCL Research Paper Declaration Form..... 185

1 Clinical Case

Mrs P is a 73-year-old female who presents to the Ambulatory Care service with four days of productive cough and reduced oral intake. In addition, her son has noted increased confusion during this period. She has a past medical history of hypertension, hyperlipidaemia and chronic kidney disease. Before this illness, she was independent with activities of daily living, mobilising outdoors without aids and living alone in a two-storey house without carers. She has recently been "slower than before" but enjoyed walking approximately half a mile daily around her local park. On examination, Mrs P is sleepy and finds it difficult to engage with the medical history, drifting in and out of the conversation. On examination, her heart rate is 98 bpm and regular, respiratory rate of 17, blood pressure 135/89 mmHg, O_2 saturations 95% on room air with a temperature 38.2°C. She is dehydrated with dry mucous membranes, jugular venous pulse seen at 1cm above the sternoclavicular junction. Auscultation of her chest reveals left basal inspiratory crackles with no wheeze. There is no clinical evidence of deep vein thrombosis or lower limb swelling. A chest radiograph demonstrates consolidative changes in the left lung base. Her blood results confirm an acute on chronic kidney injury, with a creatinine of 263 from a baseline of 120 and raised inflammatory markers (C-reactive protein 211, white cell count 16.4, neutrophils 14.2).

On the post-take ward round, the on-call medical consultant makes the following diagnoses: 1. Delirium, likely secondary to 2. Left basal community acquired pneumonia 3. Acute on chronic kidney injury. She recommends an admission to the Medicine for the Elderly team and promptly initiates treatment with intravenous antibiotics, fluids and venous thromboprophylaxis. Blood and sputum cultures are taken to optimise antibiotic therapy, including a screen for atypical pneumonia. Mrs P's son is by her bedside and is evidently concerned by his mother's current state. He "has never seen her this confused and ill-looking". He asks how bad this confusion will get? How long will it last for? How serious is this illness in the long-term and will there will be lasting effects: how will this affect her memory? Will this affect her life expectancy? The medical team acknowledges that while delirium

during acute illness is known to increase the risk of mortality and worsening cognition in the medium to long-term, we are currently unable to estimate how severe or long her delirium will last, and there is considerable variability in these outcomes. The doctors inform him that his mother's specific prognosis is unclear.

This difficult conversation is common across medical wards for older people worldwide: our understanding of adverse outcomes in older patients following acute illness is limited. While cognitive decline (and certainly mortality) can be robustly ascertained as outcomes, deterioration of motor function, increased care needs and longer length of stay are difficult to predict as specific quantities due to varying contributions of non-organic social, psychological, financial and healthcare organisational factors. Although many population-based associations have been demonstrated between risk factors and poor outcomes, such as cognitive decline and mortality risk, translation to individualised prognostication remains challenging due to heterogeneity of patients' baselines and difficulty in quantifying the acute precipitating illness.

This thesis will use two complementary techniques: population epidemiology and machine learning. These approaches can describe and quantify the relationships between baseline and outcome following an acute illness. I will focus on delirium and mortality risk, the common outcomes of patients with reduced cognitive and functional baselines. I aim to improve understanding of interacting relationships between multi-morbidities and acute illness, highlighting the most likely culprit mechanisms that could be targeted for future diagnosis and treatments. More immediately, this work contributes to a prognostication tool that could be deployed to add a novel dimension to informed patient care.

2 Completed publications to date within PhD

1. **Tsui A**; Yeo N; Searle S, et al & Davis D. Extremes of baseline cognitive function determine the severity of delirium: a population study. *Brain* 2023; 146(5) 2132-41

2. **Tsui A**, Tudosiu P-D, Brudfors M, Jha A, Cardoso J, Ourselin S, Ashburner J, Rees G, Davis D & Nachev P. Predicting mortality in acutely hospitalised older patients. *BMC Med* 2023; 21(1):10

3. **Tsui A**, Samuel Searle S, Bowden H et al. & Davis D. The impact of baseline cognition and delirium on long-term cognitive impairment and mortality: the Delirium and Population Health Informatics Cohort. *Lancet Healthy Longevity* 2022 April; 3(4): e232-e41
4. Goodyer E, Mah J, Rangan A et al. and Davis D, **Tsui A**. The relative impact of socioeconomic position and frailty varies by population setting. Ageing Medicine 2022; 3(1):10-16

5. Whitby J, Nitchingham A, Caplan G, Davis D, **Tsui A**. Persistent delirium in older hospital patients: an updated systematic review and meta-analysis. *Delirium* 2022**.**

6. Chitalu P, **Tsui A,** Searle S, Davis D. Life-Space, Frailty and Health-Related Quality of Life. *BMC Geriatrics* 2022; 22:646

7. Chalmers LA, Searle SD, Whitby J, **Tsui A**, Davis D. Do specific delirium aetiologies have different associations with death? A longitudinal cohort of hospitalised patients. *Eur Geriatr Med.* 2021 Aug;12(4):787-791

8. **Tsui A.**, Richards, M., Singh-Manoux, A., Udeh-Momoh, C., Davis, D. Longitudinal associations between diurnal cortisol variation and later life cognitive impairment. *Neurology*. 2020 Jan; 94(2) e133-141

9. Davis D, Searle SD, **Tsui A**. The Scottish Intercollegiate Guidelines Network: risk reduction and management of delirium. *Age Ageing*. 2019 Jul 1;48(4):485-488

10. **Tsui A**, Richards M, Davis D. Systemic inflammation and modifiable risk factors for cognitive impairment in older persons: Findings from a British birth cohort. *Aging Med*. 2018 Dec;1(3):243-248

11. **Tsui A**, Davis D. Systemic inflammation and causal risk for Alzheimer's dementia: Possibilities and limitations of a Mendelian randomization approach. *Aging Med*. 2018 Dec;1(3):249-253

3 Introduction

Chapter Outline

- Clinical importance of delirium: epidemiology and adverse outcomes
- Delirium diagnosis and definition: current standards and ongoing challenges
- Delirium superimposed on dementia
- Management of delirium
- Potential pathophysiological mechanisms
- Aims, objectives and hypotheses of this thesis

Delirium is an acute confusional state, representing the consequence of cerebral decompensation following a physiological stressor. The clinical construct of delirium as defined by the DSM-5 criteria is characterised by: (i) disturbances in attention; (ii) a change from baseline attention and awareness over a short period of time, usually hours to days, fluctuating in severity during the course of a day; (iii) an additional disturbance in another cognition domains (e.g. memory deficit, disorientation, language, visuospatial ability, or perception); (iv) not explained by a pre-existing, established or evolving neurocognitive disorder, or in the context of a reduced arousal state such as coma; and (v) resulting from a direct physiological input from a medical condition. Common across all healthcare settings, delirium is particularly prevalent among older people and has been associated with significant adverse outcomes, including long-term cognitive decline and increased mortality risk. Its aetiologies are broad and current management approaches predominantly involve treatment of the underlying cause.

This introductory chapter will offer an overview of delirium as a clinical syndrome. First, I will explain the concept of delirium as a disequilibrium between physiological insult and cognitive reserve. Second, I will illustrate how common delirium is by describing the epidemiology of delirium in older patients. Third, I will review the current delirium diagnostic tools and ongoing challenges in detecting delirium in clinical practice. Fourth, I will illustrate the broad range of aetiologies that may lead

to delirium and current management strategies. Lastly, I will outline adverse effects associated with delirium in population studies.

Delirium research is an emerging and rapidly growing field. In recent years, we have gained insights from multiple disciplines, including geriatric medicine, old age psychiatry, critical care, neurologists, epidemiologists and many others. This thesis will focus on inpatient medical delirium observed in older people. There are many delirium subtypes that although related, present with specific phenomenologies, aetiologies and prognoses, and hence is regarded as outside the scope of this thesis. For example, delirium tremens, a specific syndrome of neuropsychiatric and physical symptoms as a result of alcohol withdrawal - it is reasonable to anticipate different recovery trajectories for this specific cohort. While I will provide a brief summary, delirium prevention, specific management and delirium mechanisms are separate fields in their own right and are not considered in detail.

3.1 Clinical Importance of Delirium

This section aims to illustrate the importance of recognising and treating delirium in clinical practice. Delirium is common, affecting a quarter of older inpatients at any one time and has profound adverse sequelae (Gibb, Seeley et al. 2020). Consequently, delirium should represent a priority area across the board, from clinical practice, research, education and policy.

3.1.1 Epidemiology of delirium

Delirium is common across all healthcare settings and particularly among older people. The prevalence of delirium has been estimated to be up to 17% of all older people presenting to the emergency department, 31% of inpatients in general medical and geriatric wards and up to 60% in frailer elderly inpatients (Maldonado 2017) . Even higher prevalence is evident in escalated care settings such as intensive care units. One study found up to 87% of patients in US critical care units developed delirium (Ely, Margolin et al. 2001). In the community, approximately 1% of older people are thought to have prevalent delirium – though this may be an

underestimate, given sub-optimal detection methods in population settings (Davis, Kreisel et al. 2013).

3.1.2 Adverse Outcomes Associated with Delirium

3.1.2.1 Mortality

Delirium is associated with significant adverse outcomes, including increased risk of in-hospital (Naksuk, Thongprayoon et al. 2017) and post-discharge mortality (Witlox, Eurelings et al. 2010). In a systematic review including over 50 studies with a mean follow-up of 22.7 months, mortality risk was 38% during their admission for patients who developed delirium, compared with 28% for those who did not (Witlox, Eurelings et al. 2010). All included studies were adjusted for age, baseline dementia status and comorbid illnesses or illness severity (where these were measured). However, whether mortality associations are monotonic across the range of people affected by delirium is unclear. Perhaps surprisingly, delirium has been reported to have stronger associations with mortality in less frail individuals (Dani, Owen et al. 2018).

3.1.2.2 Cognitive Impairment

Dementia is a major risk factor for delirium and conversely, delirium is associated with increased incident dementia. Long-term cognitive sequalae have been demonstrated post-delirium compared with patients without delirium across healthcare settings, at 18 months follow-up after intensive care unit admission (Pandharipande, Girard et al. 2013) and five years follow-up after cardiac surgery (Newman, Grocott et al. 2001). Patients from a longitudinal birth cohort study demonstrated 1.7 points poorer scores on Addenbrookes Cognitive Examination (95% CI 0.1 to 3.2) if they have self-reported features suggestive of delirium in the ten years before cognitive testing, independently of baseline cognition and risk factors associated with Alzheimer's disease (Tsui, Kuh et al. 2018). In a key population study, an odds ratio of 8.7 (95% CI 2.1 - 35) was evident in the oldest-old, resulting in up to one additional MMSE point deterioration per year (95% CI 0.11 – 1.89) (Davis, Muniz Terrera et al. 2012). Taken together, studies with longitudinal cognitive measures have shown increased rates of decline after delirium episodes

compared with those without delirium, across an age range from as young as early 50s to oldest-old populations (Gross, Jones et al. 2012, Davis, Barnes et al. 2014).

Additional evidence suggests that delirium may accelerate underlying cognitive decline independently of classical dementia pathologies such as amyloid plaques, neurofibrillary tangles, vascular lesions and Lewy bodies. In a multi-centre cross-cohort neuropathology study involving 987 individuals from population samples, the decline attributable to delirium was beyond that anticipated for delirium or dementia neuropathology alone, suggesting that delirium may also act multiplicatively with classical dementia pathologies to accelerate cognitive decline (Davis, Muniz-Terrera et al. 2017). In imaging studies in critical care patients, longer delirium duration has been associated with lower total brain, superior frontal lobes and hippocampal volumes (Gunther, Morandi et al. 2012) while patients with delirium had persistently reduced fractional anisotropy on diffusion tensor imaging in the internal capsule and corpus callosum (Morandi, Rogers et al. 2012).

3.1.3 Long-term economic costs

Delirium has significant economic and societal impacts. The financial cost of delirium has been estimated to be between £2.6 and 5.9 billion annually for approximately 130,000 Australian occurrences of delirium alone, with £2.2 billion attributable to dementia likely as a sequalae of delirium due to increased need for personal care and institutionalisation (Pezzullo, Streatfeild et al. 2019). Each hospitalised incident delirium episode costs £13,200 in the UK (Akunne, Murthy et al. 2012).

3.2 Delirium diagnosis

Despite this wide range of adverse effects, delirium remains under-detected. It is estimated that even in secondary care, only 20% (Collins, Blanchard et al. 2010) to 50% (Clegg, Westby et al. 2011) of delirium cases are formally diagnosed. Proportions detected in the community are likely to be even lower, despite the high incidence of delirium in residential and nursing homes where underlying cognitive impairment and dementia are more common. Many practical, historical and operational issues have contributed to the current low rate of delirium diagnoses,

despite the availability of validated delirium screening tools. This illustrates the need for novel approaches, such as the potential addition of automated methods to existing techniques, in order to improve under-detection in current clinical practice.

3.2.1 Practical issues with delirium screening tools

The ideal delirium screening tool should have high specificity and sensitivity, while being fast to implement by a variety of healthcare professionals and suitable for a broad range of healthcare settings. At present, few tools are able to satisfy all of these requirements.

The 4 'A's test (4AT) is the recommended delirium detection tool in the SIGN guidelines (SIGN 2019). Along with 'months of the year backwards' (MOTYB), and the Single Question in Delirium (SQiD), each assessment is easily applied due to their brevity, simplicity and high sensitivity. However, they lack specificity for delirium diagnosis (Hendry, Quinn et al. 2016). The Delirium Observation Screening (DOS) scale appears to be more specific and sensitive, and requires assessments over three separate shifts on three separate days. The Confusion Assessment Method (CAM) is an algorithm used to detect delirium based on separate cognitive testing (e.g. MoCA, MMSE). As such, it is limited by how long it takes to complete the formal cognitive testing, which is rarely feasible in clinical practice (Wong, Lee et al. 2018). While the CAM is widely used in both the research and clinical literature, the sensitivities and specificities of delirium detection vary significantly between studies, with sensitivities as low as 28% and 46% in intensive care and medical ward settings, respectively. A recent study showed that even with nearly 100% completion rates, the Confusion Assessment Method performed twice daily was positive in only 2% of patients, far lower than the 17% rate found when delirium was measured by psychiatric assessment in the same clinical unit (Rohatgi, Weng et al. 2019). The wide performance range is likely a result of several limitations specific to the CAM: (i) it does not specify a particular assessment of attention, augmenting inter-rater variability; (ii) it is difficult to apply to non-verbal patients. Although trialled in separate language and cultural populations, no single CAM version can be used across heterogeneous patient groups; (iii) CAM requires prior training, a similar

problem for other tools such as the Delirium Rating Scale (DRS-98) , which is specifically only validated in studies when applied by psychiatrists (Detroyer, Clement et al. 2014, van Velthuijsen, Zwakhalen et al. 2016). Lastly, although underlying dementia is a significant risk factor for delirium, only 4AT, CAM, DRS 98 and mRASS have been validated for delirium superimposed on dementia.

Other general limitations for delirium instruments include detection of subsyndromal delirium, delirium subtyping and longitudinal use to track delirium recovery. At present, the process of delirium detection remains relatively crude and clinical processes rarely extend beyond the aim of determining the presence or absence of delirium in a binary fashion. Of the major instruments used for screening acute cognitive status, only DOS (13-item) (Detroyer, Clement et al. 2014, van Velthuijsen, Zwakhalen et al. 2016), DRS- R-98 (Whittamore, Goldberg et al. 2014, Sepulveda, Franco et al. 2015), ICDSC (Gusmao-Flores, Salluh et al. 2012), MMSE (Mitchell, Shukla et al. 2014), mRASS (Morandi, McCurley et al. 2012) offer a scale of severity. However, the MMSE, not specific for delirium, is a global tool of multiple cognitive domains, the mRASS is purely a measure of arousal, while the DOS is predominantly designed to assess hyperactive delirium. Neither the 4AT nor any variant of the CAM except the CAM-S, describes delirium beyond the binary presence or absence of the syndrome. The clinical phenotype of delirium is almost always classified as hypoactive, hyperactive or mixed; only the delirium motor subtype scale (DMSS) further defines motor subtyping. Most instruments are suitable for single cross-sectional delirium detection only; ongoing monitoring over serial time-points has only been validated with the CAM-ICU (Mitasova, Kostalova et al. 2012), DOS (Schuurmans, Shortridge-Baggett et al. 2003), ICDSC (Wassenaar, Schoonhoven et al. 2019), RADAR (Voyer, Champoux et al. 2015), mRASS (Chester, Beth Harrington et al. 2012) and SQiD (McCleary and Cumming 2015).

3.2.2 Changing reference standard and validity

Differences in diagnostic accuracy among delirium screening tools may stem in part from the evolving reference standard criteria for delirium from DSM III (1980), DSM III-R (1987), DSM-IV (2000) to DSM-5 (2013). An early study exploring the validity of

DSM-III-R criteria found it to be more restrictive than DSM-III but less so than ICD-10 (Liptzin, Levkoff et al. 1991). Later studies showed the simpler criteria in DSM-IV enhanced its inclusiveness over ICD-10 (Laurila, Pitkala et al. 2004). Comparing DSM-III, DSM-III-R, DSM-IV and ICD-10 in a sample of older patients from acute wards and nursing homes, 31% met any criteria, but only 6% met all four (Laurila, Pitkala et al. 2003). As the definition became less detailed and specific, delirium prevalence increased (DSM-IV 25%; DSM-III-R 20%; DSM-III 19%; ICD-10 10%) (Sepulveda, Franco et al. 2015).

This difficulty with shifting reference standards for delirium suggests a fundamental need to develop an empirical definition adding objective features linked with prognostic outcomes. At the same time, whether a cross-sectional description of pure phenomenology as the DSM-IV definition of delirium, or a novel longitudinal term potentially encompassing other modalities such as biochemistry and neuroimaging, is more clinically useful, awaits further study.

3.2.3 Operationalising delirium definitions

In each reference criteria for delirium, changes in attention, arousal and disordered thinking have consistently been components of the delirium syndrome. However, the operationalisation and quantification of these neuropsychiatric features have proved challenging for delirium detection tools. For example, attentional impairment is recognised as the central deficit in delirium. However, different constructs for attention exist. Phasic attention refers to detection of a target stimulus, while tonic or vigilance attention involves the maintenance of attentional neural correlates on a particular continuous task (Petersen and Posner 2012). In addition, tasks involving these attentional constructs rarely exist in isolation, commonly also involving other cognitive domains such as language, working memory and executive function. Current delirium detection tools are inconsistent in which type of attention they are assessing and whether they are highlighting attentional deficits alone – tests such as backward span and MOTYB require multiple cognitive domains beyond pure attentional networks (Meagher, Leonard et al. 2015). As a result, while attentional assessments in delirium are sensitive, they are commonly not specific. Operationally,

there remains no consensus on which single or combination of attentional deficits, and to which severities as defined by scoring cut-offs in different diagnostic tools, should lead to a definition of inattention.

Conversely, standalone tests of arousal have a low sensitivity but a high specificity for detecting delirium. Although assessed distinctly from attention in the 4AT, MDAS and CAM, or as a standalone test using the Richmond Agitation Sedation Scale (RASS), arousal is generally considered as a hierarchical extension of attention, in which the patient is required to display a baseline level of attention before arousal can be meaningfully assessed (European Delirium Association & American Delirium Society 2014) (Fig 7.1). Whether arousal is particularly importantly in patients with advanced chronic cognitive impairments and dementia, and hence should be weighted accordingly during delirium detection in this cohort, is unclear. While arousal is generally preserved into the later stages of neurodegenerative conditions, it may nonetheless fluctuate in advanced dementia. This may limit its usefulness as a neurocognitive sign in this context.

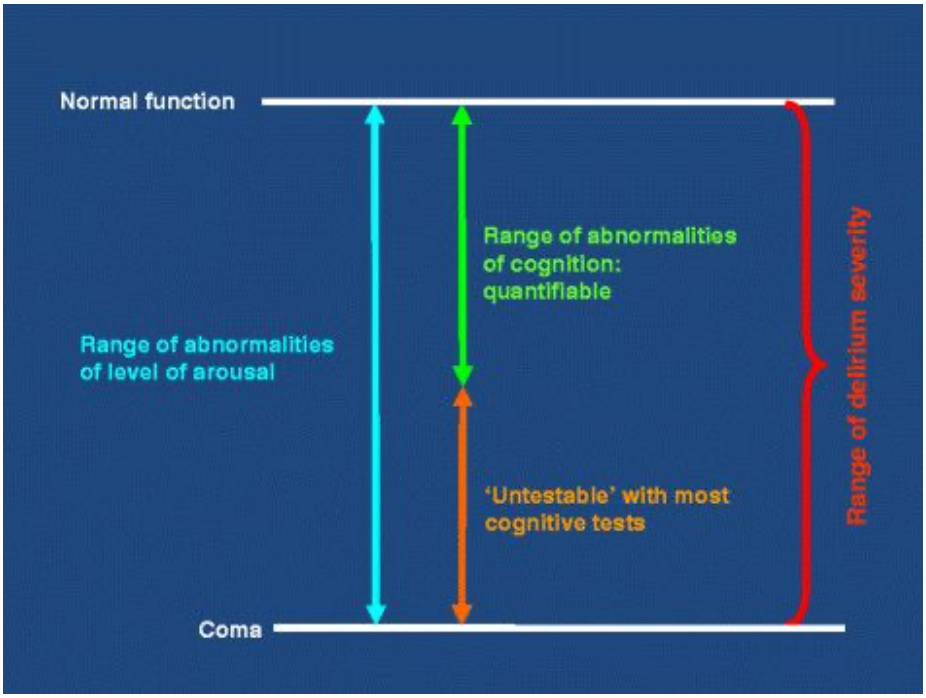


Fig 3.1: Overlap between reduced arousal states and hypoactive delirium, referenced from (European Delirium Association & American Delirium Society 2014)

Disorganised thinking remains a key criterion in the CAM but not included from DSM IV onwards. The validity of disorganised thinking as a construct remains difficult to define – it is unclear whether disorganised thinking is explained by deficits in attention, arousal and working memory or whether it should be instead considered as a distinct entity.

3.2.4 Delirium superimposed on dementia

With dementia being a significant risk factor for delirium, the concurrent presentation of delirium superimposed on dementia (DSD) is common. In general, delirium symptoms tend to overshadow dementia symptoms. However, the broad range of cognitive and neuropsychological presentations within dementia subtypes often overlap with features observed in patients with delirium.

At present, there is no consensus on whether current delirium definitions or screening tools should be modified to take into account pre-existing cognitive domain or attentional deficits in patients with dementia or mild cognitive impairment subtypes. From a temporal perspective, there is equally no clear guidance within DSM-5 or ICD-10 to accommodate differences in fluctuations or changes from an abnormal baseline in patients with preexisting dementia. A recent consensus review from the European Delirium Association has outlined a number of advances required to improve diagnostic accuracy of delirium superimposed on delirium (Morandi, Davis et al. 2017) (Box 3.1).

The consensus panel also recommended the potential role of laboratory testing and neuroimaging in DSD. The inclusion of objective measures would be novel for delirium constructs. Practically, requirements of biochemistry and imaging may also paradoxically limit delirium diagnoses in healthcare contexts where laboratory or imaging resources may not be available. However, early-stage innovations such as continuous EEG continue to rely on single-modality inputs using bespoke infrastructure unavailable in common hospital settings to diagnose delirium. There remains an urgent need to harness readily available, routine multi-modal data to improve delirium diagnostic accuracy.

Exploratory studies are ongoing to automate and make objective the diagnosis of delirium. For example, MacLullich and colleagues have developed a smartphone app to detect and monitor attentional deficits in patients with delirium (Rutter, Nouzova et al. 2018). Similarly, the feasibility of using spectral analyses of electroencephalogram (EEG) frequencies to detect delirium is being investigated using single lead EEG recordings (Numan, van den Boogaard et al. 2019).

1. Define the most appropriate type of attention in DSD in specific dementia subtypes and severity, using the most appropriate assessment
2. Quantify the specificity and sensitivity of current tests of arousal in patients with DSD, with a potential additional role of motor fluctuation evaluations
3. Define the role of clinical information from medical records, functional assessments and collateral information in DSD diagnoses.

Box 3.1: European Delirium Association recommendations for advances required to improve diagnosis of delirium superimposed on dementia

3.3 Management of Delirium

The mainstay of delirium prevention and treatment are currently non-pharmacological. There is no evidence for the use of drugs in preventing or treating delirium, except in extreme cases for sedative purposes when the patient is a danger to themselves or others.

3.3.1 Delirium prevention

Trials with prophylactic antipsychotics such as haloperidol (Girard, Exline et al. 2018) and other second-generation antipsychotics (Nikooie, Neufeld et al. 2019) to prevent delirium have not produced significantly positive findings. While dexmedetomidine may be a promising role in reducing agitation and delirium, adverse effects such as bradycardia and hypotension have been significant, with the drug yet to be used outside critical care settings (Ng, Shubash et al. 2019). Current emphasis remains on treating the acute stressor precipitating delirium and correcting this physiological insult.

3.3.2 Delirium aetiologies

Any number of pathophysiological processes can underlie the delirium syndrome. Delirium arises through the complex interplay between a patient's vulnerability to cerebral decompensation and noxious physiological stimulus, stressing the brain beyond a functional threshold. The amplitude of noxious stimuli differs for each patient. Delirium may be precipitated in those with more vulnerable brains, such as patients with older people with underlying cognitive impairment or dementia, following a small stimulus such as a change of medications or a focal infection. However, younger healthy patients would only experience delirium following large insults, such as intensive care admission, general surgery or sepsis. Delirium is a non-specific but sensitive effect of acute decompensation on a vulnerable brain, as demonstrated by a wide range of risk factors for delirium. The precipitating factors with the greatest risk factor for delirium include neurosurgery (OR 4.5), use of sedatives and hypnotics (OR 4.5), trauma (OR 3.4) and infection (OR 3.1) (confidence intervals not reported) (Inouye, Westendorp et al. 2014).

3.3.3 Potential mechanisms underlying delirium

Several potential mechanisms have been implicated in explaining how noxious stimuli occurring outside the central nervous system result in brain dysfunction. These have included neurotransmission and neuronal cellular disturbances resulting from hypoxic injury, impaired glucose metabolism, peripheral and CNS electrolyte disturbances, hypothalamic-pituitary-adrenal axis disturbance and direct effects of drugs on cholinergic and dopaminergic networks (Maldonado 2018).

Systemic and central inflammation may be relevant to delirium pathophysiology via direct CNS sterile inflammation following trauma and surgery, as well as systemic inflammation mediated by circulating cytokines, endothelial activation, blood-brain barrier dysfunction, microglial activation and vagal stimulation. Increasing evidence suggests a role for systemic inflammation in both the reversible cognitive deficits in delirium and the long-term deleterious effects of delirium, causing neuronal dysfunction and death via dissociable short and long-term IL-1 processes (Skelly, Griffin et al. 2019). In mouse models, while systemic IL-1beta produced acute working memory deficits which were protectable by application of IL-1beta antagonists, direct application of IL-1 beta resulted in hippocampal dysfunction and neuronal death over the longer term (Skelly, Griffin et al. 2019).

Although there are numerous possible delirium precipitants, and unique pathophysiological mechanisms likely correspond to each noxious stimulus, it is still possible that these processes converge upon a “common pathway” to produce a sufficiently homogeneous clinical syndrome. Recent work has focused on brain network disintegration. A functional network is a concept described through the statistical analyses of activity across brain regions during functional MRI, PET, EEG and MEG recordings, demonstrating that discrete and consistent brain regions synchronously are activated and assumed to function in coalition at rest and during tasks. A systematic review highlighted the consistent finding of functional network disruption in patients with delirium, as observed in reduced fMRI network integration and EEG connectivity, predisposed by reduced structural connectivity and poorer

structural network organisation on diffusion tensor imaging (van Montfort, van Dellen et al. 2019).

The broad heterogeneity of delirium aetiologies has been difficult to account for in clinical studies. In addition, with the high prevalence of delirium among older patients, who inherently have more complex interacting comorbidities and who commonly present with non-specific instead of organ-focused symptoms, traditional parametric models with linear assumptions are particularly poor. Inaccurate clinical ascertainment of delirium diagnoses, small-scale studies frequently involving only single modality data, and combined with a historical lack of delirium awareness, has resulted in limited understanding of the mechanisms underlying delirium and a lack of targeted pharmacological interventions for prevention and treatment. These shortcomings could be corrected by employing large-scale, multimodal, high-dimensional models using better-defined delirium diagnoses. Advances in machine learning techniques, inferential statistics and causal inferences might allow us to extend robust ground-truth prediction models into simulations. This could quantify causative effect sizes per input variable, inferring the relative contributions of data sources towards delirium diagnoses and outcomes. Such pathophysiological insights and therapeutic inferences would better target future mechanistic experimental studies and clinical trials, respectively.

3.4 Summary

Although the adverse sequalae associated with delirium are well described, the relative contributions of the patient's baseline cognitive and functional vulnerabilities, the severity of the acute physiological insult, or the inherent toxicity of the delirium episode *per se* are poorly understood. The linearity of association between baseline cognition and delirium severity is unclear. It is unknown how baseline cognition interacts with delirium exposure to result in adverse outcomes such as cognitive impairment and mortality risk. There is a need to broaden delirium ascertainment beyond a cross-sectional episode but instead employ a longitudinal approach to encompass premorbid performance while utilising validated, continuous delirium severity measures consistent with DSM-IV criteria.

Second, delirium is a model example of decompensated function in the context of vulnerable pre-existing baseline in older people. Accurately predicting outcomes in this cohort of patients with complex, interacting multi-morbidities is challenging with traditional statistical approaches. More advanced, high-dimensional techniques, such as machine learning methods, may produce better-performing predictions. This would consequently allow causal, mechanistic and therapeutic inferences, offering individualised treatments to each patient. While delirium diagnosis and prognostication using machine learning models would be a longer-term goal, the current heterogeneity of delirium definitions among available large datasets make it a suboptimal choice of outcome for proof-of-concept. For example, a large portion of current literature utilises CAM for delirium definition, which offers a binary diagnosis of a syndrome not specifically congruent with DSM-IV. In addition, delirium superimposed on dementia, affecting a larger proportion of patients with delirium, still awaits a consensus definition. As a result, I will use mortality as the target outcome to demonstrate the potential of machine learning prediction models in older patients.

Lastly, clinical experience informs us that older patients have great variance in the phenotypes of acute illness. Unsupervised machine learning techniques offer an opportunity to describe acute illness beyond common subtypes, with possible greater correlation to clinical outcomes and more appropriate stratification to treatment options, particularly in planning future clinical trials.

In this thesis, first using a population epidemiology approach, I will qualify and quantify the relationships between baseline cognition, delirium severity and burden, and subsequent long-term cognitive decline and mortality risk. Next, I will construct machine learning prediction models for mortality to demonstrate proof of concept in predicting adverse events in older patients with multi-morbidities. Third, I will seek to demonstrate inherent patterns within acute illness in older patients.

3.5 Aims and objectives

1. To describe the relationships between baseline cognition, delirium and adverse outcomes

    - Analyse linear and non-linear relationships between baseline cognition with delirium incidence and severity

    - Demonstrate associations between cumulative delirium exposure with long-term outcomes, accounting for interactions with baseline cognition.

2. To demonstrate proof-of-concept of machine learning prediction models in older patients

    - Construct supervised machine learning prediction for long-term mortality following acute hospital admission using hierarchically increasing numbers of modalities and predictive features

    - Make voxel-by-voxel comparisons of CT images between patients predicted alive and dead by the machine learning model, and compare with CT scans of ground truth to infer neuroanatomical differences between groups

3. To demonstrate patterns of acute illness in older patients

    - Construct unsupervised clustering algorithm to identify possible underlying structure within longitudinal data of baseline and acute illness in older patients

    - Demonstrate predictive utility towards adverse outcomes for identified clusters of acute illness and interrogate their component clusters features

3.6 Hypotheses

1. Poorer baseline cognition is independently associated with increased delirium incidence, delirium severity and burden

2. Greater delirium burden is independently associated with greater long-term cognitive decline and mortality risk

3. Automated supervised machine learning models can predict long-term mortality risk with a high degree of accuracy, with increasing predictive power as dimensionality and number of modalities increase

4. Inherent structure exists within acute illness presentation of older patients, when described cross-sectionally and longitudinally, with clinical utility towards adverse outcomes prediction

4 Methods

In this chapter, I will outline the statistical methods employed in this thesis and describe the datasets used. First, I will articulate the basic principles of linear and non-linear regression, survival analyses, supervised machine learning with extreme gradient-boosted trees and the basics of linear spatial parametric mapping for CT neuroimaging. Next, I will describe in further detail the two key datasets I use in this thesis: the Delirium and Health Population Informatics (DELPHIC) study and the UCLH Cognitive Status dataset, describing in detail the data collected, the pre-processing of data, as well as the strengths and limitations of both datasets.

Chapter Outline

- Statistical methods
    - Kaplan Meier and Cox proportional hazards
    - Machine learning
        - Supervised machine learning methods
        - Hyperparameter tuning
        - Feature importance
    - Statistical parametric mapping (SPM)
- Thesis cohorts
    - UCLH cognitive status cohort
    - Delirium and population health informatics (DELPHIC)

4.1 Statistical methods

4.1.1 Linear regression and regression splines

Linear regression is a basic statistical modelling technique that assumes a linear relationship between dependent and independent variables, expressed as:

$$Y = \alpha + \beta 1\,\mathrm{x1} + \beta 2\,\mathrm{x2} + \cdots . \beta k\,\mathrm{x}k + \epsilon$$

where α = intercept, β = coefficient, x = independent variable, k = number of independent variables, ϵ = error

However, linear regression can sometimes be too simplistic to describe associations between dependent and independent variables. In the case of suspected non-linear associations, the use of additional predictors raised to increasing powers, known as polynomial regression, can be used to fit data instead, expressed as:

$$Y = \alpha + \beta 1\mathrm{X} + \beta 2X^2 + \beta 3X^3 \ldots . \beta kX^k + \epsilon$$

where α = intercept, β = coefficient, x = independent variable, k = number of independent variables, ϵ = error

However, with increasing powers of predictors included, polynomial curves become prone to overfitting, poorly generalising to unseen data. In addition, polynomial regression can be highly affected by anomalous values, with outliers resulting in significant effects on the overall fit of the regression curve.

One solution is the division of the dataset into bins, delineated at points known as knots, then fitting separate regression lines within each bin, known as regression splines. While the positions of knots can be specified, it is common practice to place them uniformly across the dataset. However, the number of knots for splines must be selected. The functions fit within each bin do not need to be the same. Instead, functions can be fit piecewise using a number of linear or low-degree polynomial splines, avoiding the severe oscillatory nature of high-order polynomials. However, without further constraints on these piecewise functions, individual regressions may be discontinuous from those in adjacent bins, potentially with two curves

interpolating different values at the boundaries between two bins at the same knot. Mitigating this, constraints can be applied to smooth piecewise curves. Commonly, these specify the first and second derivatives of adjacent curves be equal.

Lastly, a characteristic of polynomial functions is the wide variability at the boundaries of the dataset, such as beyond the first and last knots. A further constraint can specify that the function must be linear beyond the first and last knots at these boundary regions. The result is known as either a natural or a restricted spline.

4.1.2 Kaplan Meier and Cox proportional hazards

Kaplan Meier estimator is a non-parametric survival function calculating the fraction of a population alive after a time t. Survival at any one time is calculated by:

$$S = \frac{(N_{alive} - N_{dead})}{N_{alive}}$$

$$where\ N\ is\ the\ number\ of\ people\ in\ the\ analyses$$

The probability of survival at a certain time is the multiplication of survival probabilities of all prior time points. Kaplan-Meier estimators right censor when a patient has been lost to follow up, withdraws from the study or continues to be alive without an event occurring. Kaplan-Meier estimators can be visualised as a plot with horizontal steps, declining with each death and with large sample size, demonstrating the true population survival probabilities. Kaplan-Meier estimators also allow survival stratification between different groups, visualised by different trajectories of decline.

Cox proportional hazard models allow simultaneous evaluation of several factors on survival. It is defined as the multiplication of two parts: first, a baseline hazard function defining the probability of survival at baseline covariate levels; and second, the effect parameters, calculated as the exponential sum of the products of covariate parameters and their weights.

$$h(t) = h_0(t)e^{(\beta 1x1 + \beta 2x2 + \cdots . + \beta ixi)}$$

where $h(t)$ = hazard at time t, h_0(t) = baseline hazard, β_i = coefficient for covariate i, x_i = feature i

An exponentiated coefficient produces a hazard ratio (HR), with a HR above one signifying increased risk of an event occurring and below one a lower risk. By fitting a Cox proportional hazard model, the coefficients generated estimate a hazard ratio of the proportional increased risk of mortality over the baseline reference level of a covariate.

4.1.3 Machine learning

Machine learning is an umbrella term for statistical methods and algorithms that imitate human learning by training automatically and iteratively until the performance of a task has been maximised. Task objectives can range from predicting a specified outcome to identifying inherent patterns in data without a prior label. All data utilised are known collectively as the dataset. In prediction tasks, the target is termed the *outcome* and any non-outcome data points are termed *features*.

Machine learning methods generally fall into four broad categories: supervised, unsupervised, semi-supervised and reinforcement learning. In supervised learning, the machine is trained with labelled outcome data. As the machine trains to produce a model as close to the ground-truth label as possible, weights assigned to original and derived features are adjusted, until performance can no longer be optimised. Supervised learning is used for tasks such as prediction or label classification. Examples include logistic and linear regression, decision trees, random forests, support vector machines and neural networks. In unsupervised learning, the machine trains on an unlabelled dataset and aims to find inherent patterns without human guidance. Typical uses include analyses for similarities and differences, such as clustering, or to reduce the number of dimensions within a dataset while retaining maximal information, such as principal component analysis. Semi-supervised learning utilises a combination of the supervised and unsupervised approaches, generally first training on labelled datasets before progressing onto training with

unlabelled data. Reinforcement learning is considered a variation of supervised learning, in which the model performs automated trial-and-error experiments. The machine is given feedback of success and failure: it is the drive towards successful outcome feedback that tailors the training of the machine, instead of an aim towards a specific target label per se.

Machine learning, regardless of the type of method, share a common system of learning. First, a dataset is formed, either by collecting real-world data or by generating simulated data. Next, the dataset must be prepared for the chosen model, being aware of the requirements of specific algorithms: random forests cannot use categorical data, which must instead be transformed into binned individual binary “dummy” variables. Some are unable to handle missing data and require prior imputation.

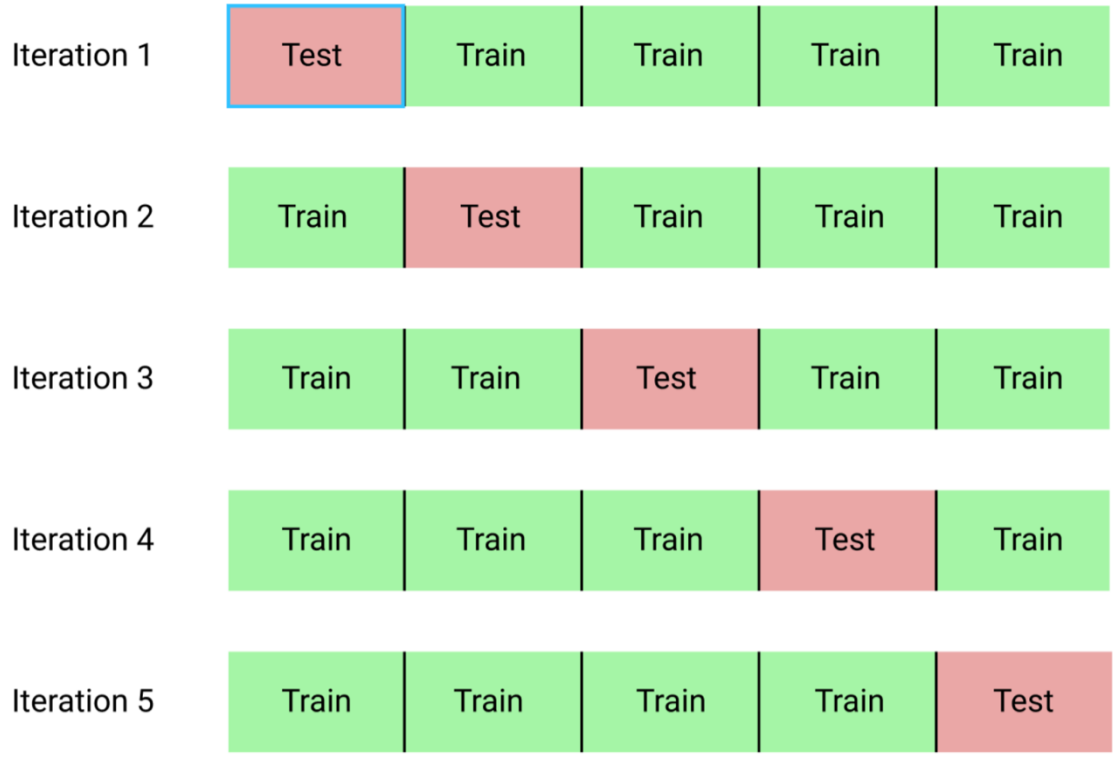
The machine is then trained, in which estimates for each feature are produced by the model in a “decision process”. In a classification task, the model produces a coefficient and weight for each feature and a predicted class for the outcome. A range of measures may be selected to measure the residual errors between predicted and ground truth outcomes, such as area under the curve receiver operator characteristic (AUCROC) or root mean squared error (RMSE). Lastly, weights apportioned to specific features are adjusted, and the model retrained. This process continues iteratively and aims to reduce residual errors, the specific metric of which can be designated. The loop continues automatically until no further error reduction can be achieved, signifying that performance cannot be further optimised, or a performance threshold has been reached.

If model has not been able to extract maximal performance from the training data, analogous to a sub-optimally fitted regression in classical epidemiology, this is termed an underfit. However, if a model has been produced that recognises only patterns or predictions specific to the training dataset but would perform poorly on previously unseen data, this is termed an overfit. The model is, therefore, poorly generalisable.

There are various approaches to assessing model generalisability: first, as in classical epidemiology, the model can be applied to an external, independent dataset. However, in machine learning, if no independent dataset is available, the concept of “hold out”, in which the dataset is partitioned from the start into a training and testing subsets, commonly in a 70%-30% split for training and testing respectively. The model would only learn from the training subset and its generalisability assessed in the testing subset. Cross-validation (CV) is used within the training subset alone. Classically, this involves dividing the training set into k subsets, k times. For each k subset, the portions of the dataset assigned for training and testing vary. At each fold, the machine learns and tests the model from different portions of the dataset, resulting in k test performance metrics, the average of which is taken as final (fig 4.1). An extreme example of K fold validation is “leave one out”, in which only one sample of the entire dataset of n samples is designated the test subset at each fold and the model is trained using n-1 features, resulting in n model evaluations, which can of course be time consuming.

K fold CV can be problematic when the dataset has an imbalance in outcome targets. For example, if a mortality prediction model uses a dataset in which 90% of participants remain alive and only 10% die, traditional K fold CV may result in unbalanced training folds, in which some folds may have no instances of death. Stratified K fold CV ensures there are the same number of outcome events across each fold, resulting in better representation of the whole population across every training subset.

Fig 4.1: Example of K fold cross validation (k=5)



In this example of five-fold cross-validation, the dataset has been split into five subsets: within each of five iterations, a different subset will be held back as the test set while the other four will be used for training (reproduced from Towards Data Science: Cross Validation Explained: Evaluating estimator performance)

4.1.4 Supervised learning methods

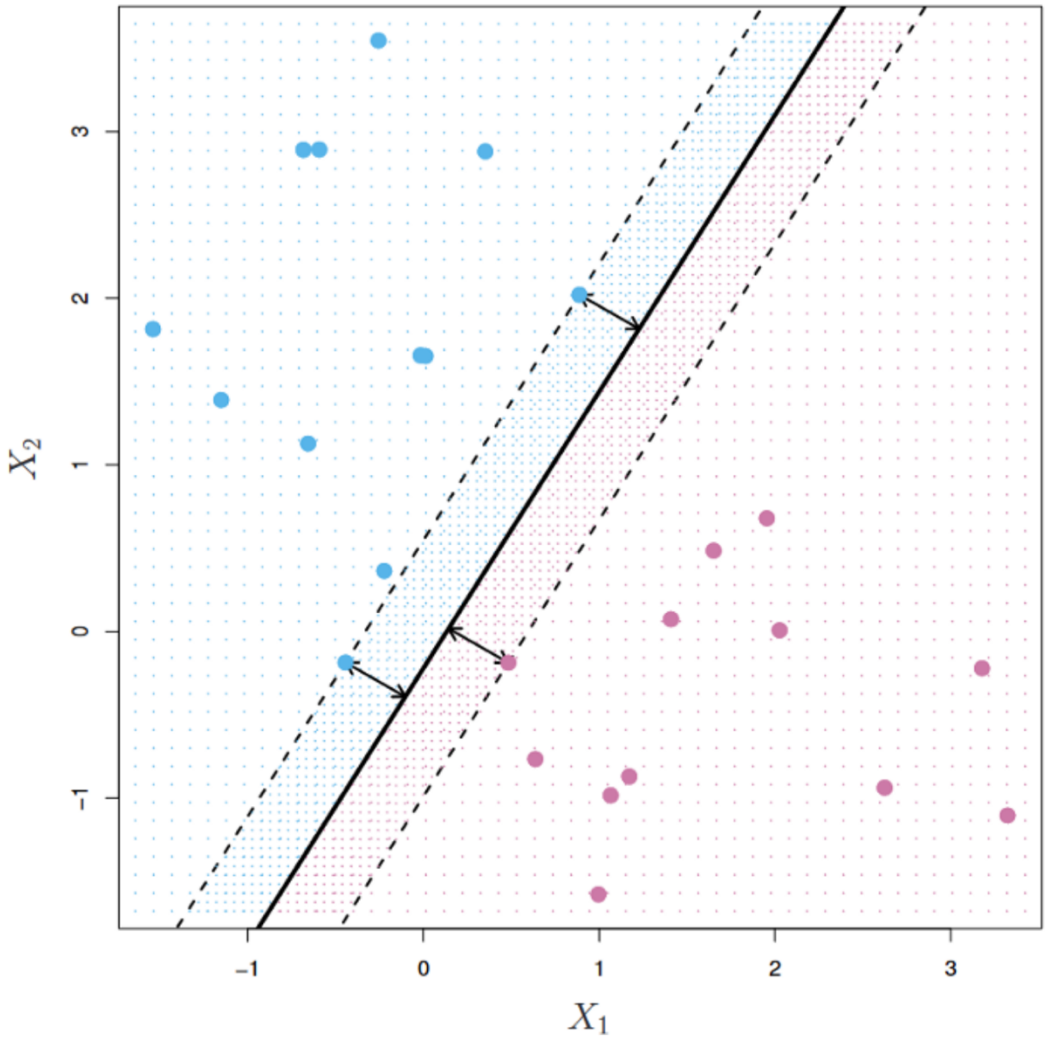
Machine learning is an automated iterative method that in some classification cases, are extensions of classical statistical methods. For example, logistic regressions, linear and non-linear regressions and margin classifiers can applied within a machine learning framework. In this section, I will describe the most commonly used supervised machine learning methods, focusing mainly on support vector machines, decision trees, random forests and boosted gradient trees. I will finish by focusing on the details and modifiable hyperparameters of XGboost, on how these measures can most optimally tuned to maximise performance.

4.1.4.1 Support vector machine

Support vector machines are an extension of maximal margin classifiers, a statistical method that aims to separate data points of different classes with a hyperplane. The dimension of a hyperplane is one less than the number of dimensions in a dataset: for example, a 2-dimensional scatterplot can be separated by a line while a 3-dimensional space can be separated by a flat plane. Datapoints closest to a proposed hyperplane are known as support vectors and the Euclidean distance from them to the hyperplane the margin (fig 4.2).

While maximal margin classifiers are excellent methods for perfect linear separation of classes, they are limited when either the hyperplane is non-linear and/or the classification is imperfect. Support vector machines overcome this limitation by projection of a dataset into a higher dimensional space, allowing the drawing of a more non-linear and flexible boundary between support vectors. To reflect similarities between data points as a matrix dot product, the supported utilisation of kernels avoids the requirement to fully compute the new higher dimensional state, significantly optimizing computational power.

Fig 4.2: Support vector machine



Support vector machine on a two-dimensional dataset, drawing a line that separates two classes (blue and red); arrows = support vector (reproduced from Towards Data Science: the complete guide to support vector machines)

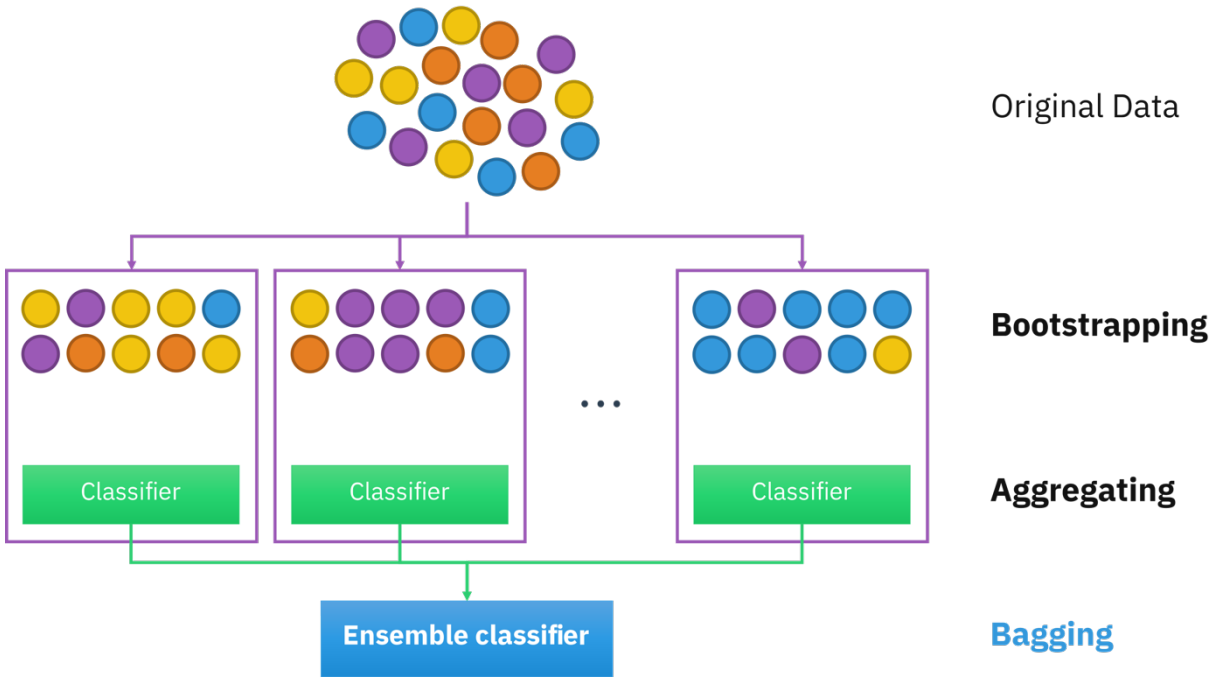
4.1.4.2 Decision trees and Random Forests

Decision trees are intuitive methods used for classification or regression tasks. They are a series of thresholded sequential questions that divide the dataset, potentially even separating into individual data points if allowed. They have significant advantages: they are easy to interpret, fast to compute and perform well in large volumes of data. However, individual trees alone are prone to overfitting with poor generalizability to unseen data. At each individual split, the algorithm is motivated only by what performs best at the local node without consideration of the whole dataset, resulting in progression iteratively towards distinctions applicable only to the training dataset.

However, poorly performing individual models can be combined into an ensemble model for improved performance, for example, by reducing variance. Ensemble models can be constructed from the same or different constituent models, known as homogenous and heterogenous ensembles, respectively.

An ensemble of minimally correlated decision trees is known as a random forest. The ensemble method used to combine individual models in random forests are known as bagging (bootstrap aggregation): random subsets of data are selected from an original dataset with replacement, creating n samples that are used to estimate the desired parameter (fig 4.3). At the same time, only random subsets of features are made available in the selected samples. The output of a random forest regressor is the average output of bootstrapped samples. In random forest classifier, a similar concept can be applied by averaging the class probabilities generated from each bootstrapped output. However, in an alternative approach known as hard voting, the random forest classifier output can also be the vote selected by the majority of bootstrapped samples. As the model continues to iteratively train, the weights attached to each feature and their contribution to the final model will be adjusted, protecting against individual errors and producing a more generalisable model less prone to overfitting.

Fig 4.3: Schematic representation of bagging (bootstrap aggregating)



Original data is sampled with replacement n times, creatining n samples that produce n classifier outcomes, the ensemble classifier is the calculated by majority vote or average of class probabilities. (Reproduced from https://commons.wikimedia.org/w/index.php?curid=85888768)

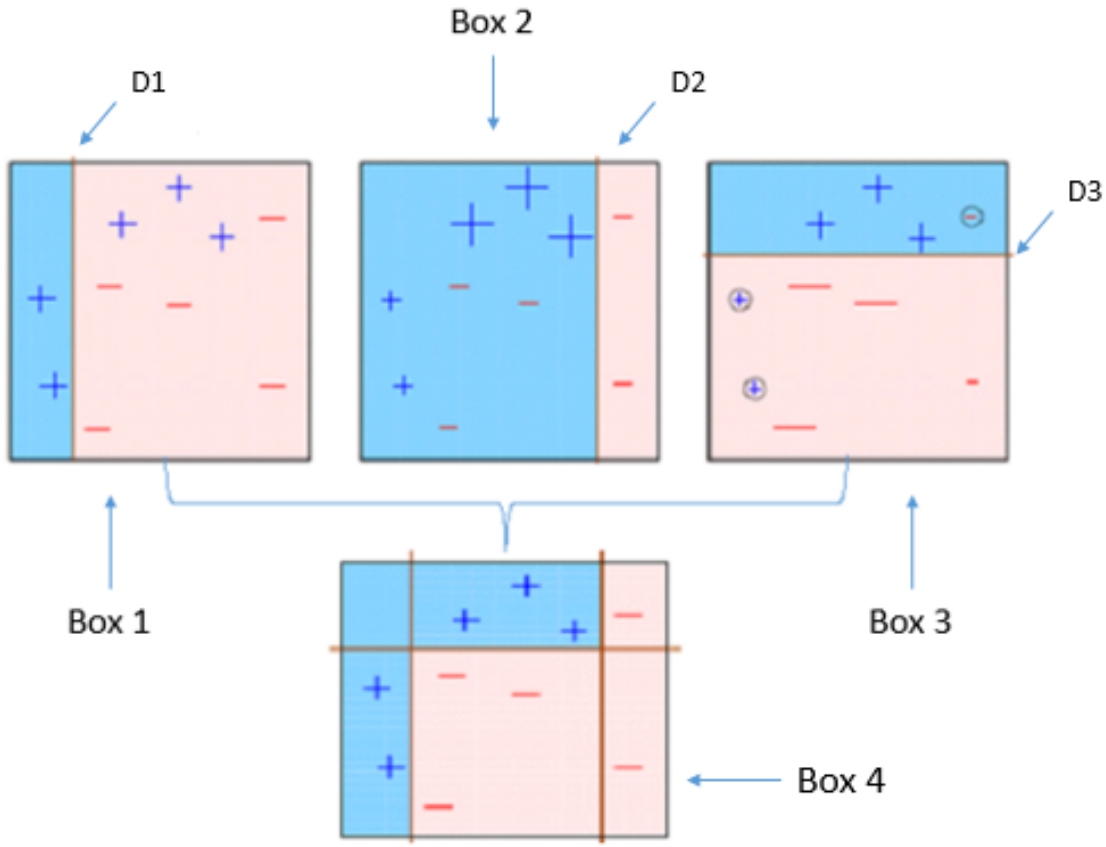
4.1.4.3 Boosting

Boosting further improves performance of ensemble models by sequentially focusing on weak learners. For example, decision trees that on their own would perform minimally better than random chance are modified to become better learners, benefitting the model as a whole. This process is sequential and iterative. The earliest form of boosting was Adaboost: initially, all dataset observations are assigned an equivalent weight towards the final model. Each weight is progressively adjusted as newer weak learners are added to the model, without changing existing previous learners, with increased emphasis on previously incorrectly misclassified observations (fig 4.4).

Refinement of Adaboost is achieved by using a numerical measure of error, known as a loss function. The next sequential learner to be added to the model can be programmed with the aims that it reduces this error, in the direction in which the rate of loss would be fastest, mathematically described as the greatest negative derivative with respect to the last learner's loss function - negative gradient descent. In this most optimal direction, the size of each descent is governed by a learning rate. After the similar iterative and sequential learning process, the local minimum will be reached, when new learners can no longer "step" negatively in relation to the previous loss function, stopping the gradient-boosted tree learning.

Extreme gradient-boost (XGboost) is another boosted ensemble model and, at present, one of the most efficient implementations of gradient-boosted trees. Compared to the characteristics of the first gradient-boosted trees, as described above, XGboost has several advantages. First, it has significantly faster computational speed due to its ability to utilise parallelisation. While weak learners are linearly sequentially added, all features still require searching for the best outcome split at each node. XGboost searches all features in the dataset. Each instance is balanced regardless of the tree being built, generates a distribution of instances, and applies them in parallel as most appropriate across all trees within the ensemble. Instead of repeating the search for each feature at every node, XGboost searches by feature and applies them to each node.

Fig 4.4: Schematic representation of boosting using Adaboost



Machine increases emphasis with each iteration from box 1 to 3 towards observations previously misclassified, to produce box 4 = an accumulation of previous weak learners (reproduced from Analytics Vidhya: quick introduction to boosting algorithms in machine learning)

Second, there are options to avoid overfitting the prediction model using L1 (Lasso) and L2 (Ridge) regularisation, known as alpha and lambda in XGboost respectively. Lasso penalises overfitting by adding a regularisation term to each loss function, calculated as the sum of the absolute value of coefficients, potentially altogether removing extreme outliers. L2 penalises the sum of the squares of weights. In addition, XGboost allows a Langrangian multiplier called gamma, which is in essence, a pseudo-regularisation parameter that penalises model complexity: the higher the gamma, the greater the regularisation.

XGboost can handle missing values. As a result, prior imputation, as required for other machine learning methods, is not required. Fourth, in addition to the regularisation terms above, XGboost is extremely flexible, with a number of modifiable hyperparameters that can be modified, mostly to avoid overfitting. The most commonly used hyperparameters include:

- **max_depth**: The maximum depth of each tree within the ensemble, with deeper trees potentially performing better in training but with greater risk of overfitting. Trees either stop at the prescribed maximum depth or whenever a split results in a negative loss.
- **learning_rate**: The size of each additional iterative step as the model is training and being optimised. The lower the rate, the slower the computation. Conversely, the faster the computation but potentially at the risk of not being able to reach an optimum.
- **n_estimators**: The number of trees allowed in the final ensemble.
- **colsample_bytree**: The number of features or columns allowed to be used by each tree, represented as a fraction of the total dataset.
- **subsample**: The number of observations allowed to be used by each tree, represented as a fracture of the total dataset.

Lastly, XGB can also be applied for a survival model, known as accelerated time failure (XGB aft). Transformed from a linear regression survival model to a decision tree ensemble, accelerated time failure is expressed in the formula:

$$\ln Y = \mathcal{T}(\mathbf{x}) + \sigma Z$$

where $\mathcal{T}(\mathbf{x})$ is the output of a decision tree, Z is a random variable and σ the scaling factor for Z.

The prediction metric for survival is Harrell's C-index. In summary, this measures concordance between the risk of mortality, calculated by the model, and time to participant death, if observed. For example, for patients i and j with risk N and time to death T: if both patients are not censored, if Ni>Nj, concordance is defined as achieved if Ti < Tj and if Ni > Nj. The converse scenario, in which either operator

isreversed, would be discordant. However, if one patient was censored, for example, if Tj > Ti and Tj is censored, and Ni >Nj, concordance is also achieved.

Over the dataset, Harrell's C index is a calculation of:

$$\frac{\text{Number of concordant pairs}}{(\text{Total number of concordant and nonconcordant pairs})}$$

4.1.5 Hyperparameter tuning

Optimisation of hyperparameters is essential towards achieving the best model performances, particularly in flexible models such as XGboost with numerous combinations of modifiable parameters. Hyperparameter optimisation has been traditionally achieved using techniques known as *grid* and *random searches*: in a grid search, a prespecified set of candidate hyperparameters are defined and every combination of them used to train the model. The hyperparameter set with the best performance is used. Random search is an abbreviated version in which only random selections of hyperparameters from a pre-specific list are used, decreasing computational time but increasing the risk that the most optimal combination may not be found.

Generating each model from scratch using individual hyperparameter combinations is computationally and time intensive. An alternative approach is Bayesian optimisation, which instead uses probability statistics. In summary, for every function requiring optimising (in the case of XGboost hyperparameters tuning, the function to optimise is model performance and input data point the hyperparameters values), a surrogate function is fit using limited data points, obtained from sampling. Next, an acquisition function for each data feature on the surrogate function, assigning the probability of each being a good candidate for optimising the model performance. Changes to the model are concentrated in regions most probable for optimising performance, applied, and then used as updated prior surrogate function probabilities. Bayesian optimisation do not rebuild entire ensemble trees for every hyperparameter combination and hence avoids consequent expensive computation.

4.1.6 Feature importance

Feature importance provides a score of contribution from each feature to the final accuracy of the predictive model. This is calculated by the performance improvement resulting by a split point of a clinical feature (“gain”), weighted by the number of observations the node at which the split occurs (“cover”). Feature importance is calculated for the final model by averaging the gain-cover metric across all trees within the ensemble.

4.2 Clustering

Unsupervised machine learning techniques can also be applied to demonstrate patterns, linear and non-linear, within high-dimensional datasets, agnostic to prior assumptions. Clinically, clustering may group similar patients into phenotyping subtypes, with potentially relevant similar trajectories, prognoses and management approaches. From a research perspective, patients may be more optimally stratified for clinical trials.

Clustering techniques broadly utilise measures of similarity between data point. In this chapter, I will first describe how best to deal with mixed datasets consisting of continuous, ordinal, categorical and binary data, frequently found in clinical records. Next, I will articulate the principles of t-stochastic neighbour embedding (TSNE), which allows condensation of non-linear, high-dimensional data into a lower-dimensional representation to allow easier visual delineation of possible clusters. Last, I will describe candidate clustering methods employed to define clusters.

4.2.1 Gower matrix

A matrix of Gower’s distances offers a method of dealing with datasets composed of mixed types of continuous and categorical variables, particularly when a hierarchical order cannot be assumed in a clinical feature (for example, male and female sex). As a result, Gower’s distance is a useful measure of dissimilarity suitable for quantitative, ordinal, nominal and continuous variables. Gower’s distance is a weighted average of dissimilarities among pairs of data points. A Gower’s distance of 0 represents equality while 1 is maximal dissimilarity.

When variables are numeric and continuous, the partial similarity between points are calculated as:

$$s_j(x_1, x_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j}$$

Where S_j = similarity measure, which R represents the range of the feature j.

When variables are non-numeric with multiple non-hierarchical levels (for example, occupation), the partial similarity index is calculated using the Dice distance. All numeric inputs are converted into a presence/ absence matrix. Dice distance works on the basis of whether the number of present dimensions in two values are equal or not:

$$\frac{\text{Number of non equal dimensions}}{(\text{Number of dimensional in which presence of both values are true} + \text{number of non zero dimensions})}$$

An averaged sum of similarity indices, calculated using numeric methods or dice distances, is used to generate a matrix of dissimilarity Gower's distances, known as a Gower's matrix.

$$D_{Gower}(x_1, x_2) = 1 - \left(\frac{1}{p} \sum_{j=1}^{p} s_j(x_1, x_2) \right)$$

where Sj (X1, X2) is the similarity index between two features of two observations

4.2.2 T-stochastic neighbour embedding (TSNE)

In these analyses, TSNE allows easy visualisation of high-dimensional data, such as downsampling a large complex dataset into a two-dimensional manifold. TSNE is an unsupervised technique for data exploration and visualisation of underlying data structure developed in 2008 by Van der Maatens and Hinton (van der Maaten and Hinton 2008). Unlike principal component analysis, TSNE often improves efficiency of downsampling and cluster visualisation in complex datasets that have non-linear manifold structures, due to its ability to preserve only small local distances.

In summary, the aim of TSNE is to minimise differences in distributions of similarity indices. First, TSNE fits a Gaussian distribution for a single point in high-dimensional space, calculates a set of centred and normalised distances to all other points within a sphere of influence and finally a set of probabilities. This is iteratively repeated with the Gaussian distribution fitted for all points, distances calculated for remaining points within spheres of influence and multiple sets of probabilities. The probabilities between sets of points are compared to generate a matrix of similarities. The perplexity of TSNE analysis determines the size of the sphere of influence for contributing to the calculation of variance, and effectively, the number of nearest neighbours deemed significant (the degree of local influence vs global influence). There is no statistical test for the most optimal perplexity value: the authors of the technique recommended “typical values for the perplexity range between 5 and 50”, with “larger or dense datasets requiring a larger perplexity”. There is an intuitive acceptance that perplexity should be aimed at approximately $n^{0.5}$, where n signify the number of data point (Oskolkov 2019).

The same process is repeated using a student’s T distribution with 1 degree of freedom, also known as a Cauchy distribution, which has thicker tails than a normal distribution and is better for modelling datasets with points far apart in Euclidean distances. A second set of probabilities is created in low-dimensional space. Lastly, a gradient descent approach is iteratively used to minimise the degree of divergence between the initial Gaussian and subsequent Cauchy distributions by optimising the

Kullback-Leibler divergence, resulting in a final lower-dimensional manifold representation than the original dataset.

4.2.3 Clustering methods

Clustering is an unsupervised statistical process that identifies inherent pattern structures within a dataset, resulting in grouping data points with similar characteristics. The most common clustering techniques are agglomerative and centroid based, such as hierarchical clustering and K means clustering. In addition, clustering can also include divisive, distributional (in which similar clusters are assumed to belong to similar distributions) or density-based (in which similar data points are considered to be located within areas of similar local densities).

4.2.3.1 Hierarchical Clustering

In hierarchical clustering, the number of clusters is not a priori defined. Initially, each individual data point is assigned as one cluster. Next, the two nearest points are merged into a single cluster. This process continues iteratively until the whole data set is grouped into a single cluster. A dendrogram is drawn, visualising the distance in data space in the y axis. Although there is no specific rule guiding the optimal number of clusters, it is generally accepted that the number of clusters traversed by a horizontal line at the level of a dendrogram with the greatest y axis height (and hence data space distance) without intersecting another cluster is also the most optimal.

The definition of “distance” can be determined differently, affecting the clusters produced by hierarchical clustering. Common definitions of distance between two points “a” and “b” include:

- Euclidean distance: $\|a-b\|_2 = \sqrt{(\Sigma(a_i-b_i))}$
- Squared Euclidean distance: $\|a-b\|_2^2 = \Sigma((a_i-b_i)^2)$
- Manhattan distance: $\|a-b\|_1 = \Sigma|a_i-b_i|$
- Maximum distance: $\|a-b\|_{infinity} = \max_i |a_i-b_i|$

- Mahalanobis distance: $\sqrt{((a-b)^T S^{-1} (-b))}$ where s is a covariance matrix

A little used but simple opposite of hierarchical clustering is divisive clustering, in which instead of every point being initially clustered into individual clusters, the entire dataset is designated as one cluster and subsequently "divided".

4.2.3.2 K means clustering

K means is a centroid method of clustering that requires a priori determination of the number of clusters to be formed, suggesting some degree of domain knowledge is required. The classical form of K means randomly selects points within the data space and each data point is assigned the cluster "class" of the nearest point. Next, for all the points within each cluster "class", a new centroid is calculated, after which each data point is reassigned to the class of the nearest new centroid. This iterative process continues until the stopping criteria has been satisfied, usually when centroids of newly formed clusters do not change or the maximum number of iterations.

An advance on K means is to optimise the initiation process, in which instead of random allocation of the first points in space, K++ selects a random point as the first K. The squared distance of all other data points to this first selected point is calculated and the second selected K is the data point furthest away from the first K. Next, the squared distances of all non-centroid points are calculated to the nearest centroid and the points furthest away from any of the previous two K points is selected. This process iteratively continues until the pre-determined K number of centroids have been defined for initiation, after which K means continue as per the classical model.

There is no specific number of optimal K, but the most commonly used method to determine the appropriateness of K is to calculate the sum of all distances within a cluster to the centroid, known as inertia. Alternatively, the intercluster distance can be added as a further measure of data compactness, known as the Dunn index, calculated as:

$$\frac{\textbf{Min (intercluster distance)}}{\textbf{Max (intracluster distance)}}$$

A plot of inertia or Dunn index, which are measures of intracluster distances to be minimised or cluster compactness to be maximised respectively, against the number of pre-determined clusters (K), demonstrates an “elbow” inflection point, the number of clusters at which is then selected as the most appropriate K.

4.2.3.3 K means vs Hierarchical Clustering

K means differentiates from hierarchical clustering in several ways: first, using random initiation instead of K++, K means may produce similar but not exactly the same outcomes. Second, the setting of K, in the clustering model requires an a priori domain knowledge to use K means appropriately. Lastly, K means is inherently more suited to lower dimensional or spherical manifolds than hierarchical clustering, which is more tolerant to higher dimensional shapes.

4.3 Statistical Parametric Mapping (SPM)

4.3.1 What is SPM?

Statistical parametric mapping is a process and program (Friston & Ashburner) that is a voxel-based approach to compare morphometric or functional brain differences between groups of subjects or within subjects over time, using classical statistical inference. The program also offers a pipeline to pre-process neuroimaging before analyses. In this thesis, SPM will only be used for structural volume-based analyses and my summary of its functions will not extend to time series functional neuroimaging.

4.3.2 Pre-processing

Pre-processing of images is required to allow analyses between comparable images. First, realignment of images is required: movement discrepancies between scans are inevitable, even by a few millimetres, even in cooperative patients. This is performed by rigid body transformation, in which the six parameters are produced to reflect movements in each of the three rigid motions (reflection, rotation, translation). Next,

images are spatially normalised against a standardised brain template. For each series of images, the mean is taken to calculate the warping function mapping them onto a brain template. In SPM, this “normal brain” was following a series of healthy controls at the Montreal Neurological Institute (305 participants, mean age 23.4 years +/- 4.1 years, 66 females 239 males), hence known as the MNI space. Third, the image is smoothed to reduce noise. It is assumed that while noise occurs randomly from voxel to voxel, true signals likely extend across several adjacent voxel. As a result, a Gaussian kernel is applied, in which each voxel is replaced by a value of itself with contributions from values of its nearest spatial neighbours.

4.3.3 Statistical Parametric Mapping

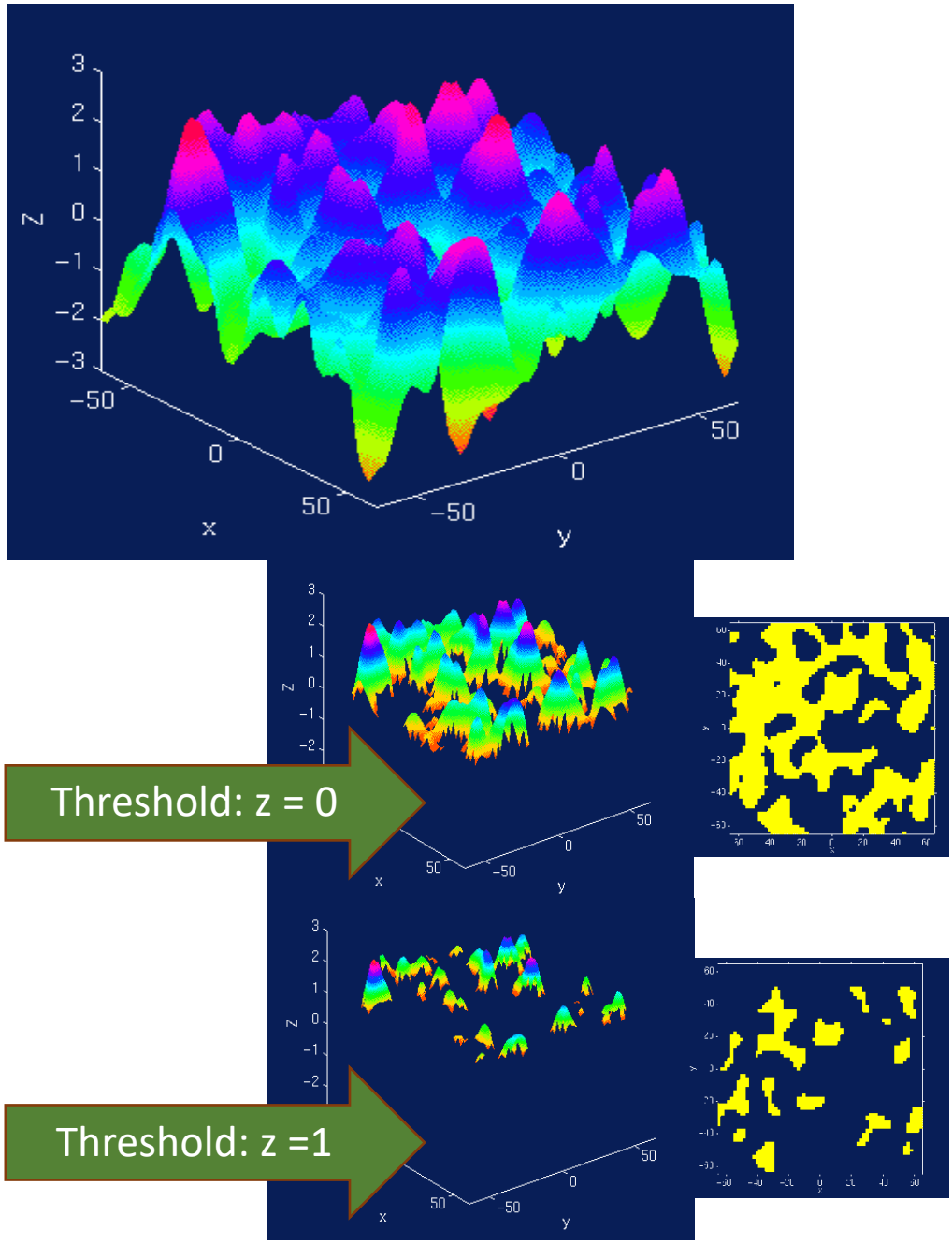
Voxels are assumed to have values and variances of T and F distributions, respectively. Analyses are performed using the generalized linear model on a voxel-by-voxel basis with parametric assumptions.

The large number of simultaneous analyses, across individual voxels and across imaging contrasts, requires correction for potential “family wide” type 1 errors, in which the null hypothesis is incorrectly rejected. However, assumption of no spatial correlation by using Bonferroni correction would be too conservative, as this would assume every voxel would be independent from every other voxel, even if adjacent in three-dimensional space, incorrectly leading to type 2 errors. A solution is required that takes into spatial correlation, only rejecting the null hypothesis if the more local or whole observed volume of voxels is unlikely to have arisen from a null distribution.

SPM utilises random field theory, a recent body of mathematics that defines the theoretical smoothness of a statistical map. In short, a random field is list of random numbers mapped onto a space of n dimensions, with an underlying assumption that spatial correlation exists even within this “random field”, in which numbers in adjacent Euclidean indices differ less those far apart. The “full width can estimate the smoothness of an image at half maximum” (FWHM) of the applied smoothing kernel. The number of voxel blocks in the image equal to the width of the smoothing kernel can be estimated, a correlate of the resolution of spatial correlation known as the resolution element (Resel). The smoothness of the image and the random field can

be compared. Any elements above a Euler characteristic, an estimate of the probability of an image achieving an assigned standardised score not by chance, can thus be defined as significant and likely arising from a localised spatial difference (fig 4.5).

Fig 4.5: A demonstration of thresholding by Euler characteristic to determine significance with random field theory.



An illustration of the three dimensional space (top) with Euler characteristic thresholded at z = 0 (middle panel), resulting in greater areas of deemed significance in three (left: x, y, z) and two (right: x, y) dimensional spaces respectively; more aggresive thresholding at z = 1 (bottom panel) results in fewer spatial structures deemed significant in three (left: x, y, z) and two (right: x, y) dimensional spaces respectively (reproduced from SPM Random Field Theory lecture notes)

4.4 Datasets

4.4.1 Delirium and Population Health Informatics Cohort (DELPHIC) study

4.4.1.1 Cohort

The Delirium and Population Health Informatics Cohort (DELPHIC) is a prospective longitudinal study of older adults aged ≥70 years in the London Borough of Camden, across primary, secondary, intermediate, community mental health and social care, to obtain complete longitudinal data on health and functional status from baseline and inpatient acute illness (Davis, Richardson et al. 2018). The National Health Service provides over 95% of healthcare in the borough for a population of approximately 260,000 residents, comprising a single primary care system and two acute hospitals, University College London Hospital (UCLH) and Royal Free Hospital (RFH).

Eligible participants were aged ≥70 years and registered with a Camden-based general practitioner. Those with severe hearing impairment, aphasia, unable to speak English sufficiently to undertake any basic cognitive assessment, or in the terminal phase of illness were excluded from the study. Letters were sent from general practice list to invite individuals to join the study, augmented by direct recruitment of patients from memory clinics and those recently discharged from secondary care specialists such as old age psychiatrists, community and hospital geriatricians. The ratio of recruited participants from GP lists, memory clinics and directly from inpatients within secondary care was approximately 8:1:1. All individuals, or their nominated proxies, gave consent to participate.

4.4.1.2 Baseline Assessments

DELPHIC was specifically designed to capture information not routinely collected in cohort studies but are particularly important for the understanding of health outcomes in older patients, such as data on quality of life, visual and hearing impairment, nutritional state, falls, continence, activities of daily living and cognition, with more than 700 raw baseline variables per participant.

Assessments were performed at baseline. While the majority took place by telephone, home visits were arranged for participants with significant hearing impairments. These assessments comprised of the following:

- Consent for involvement in DELPHIC, specifically including hospital assessments in the event of acute illness (particularly if capacity was impaired through delirium or dementia at subsequent contacts); record linkage of electronic health data in primary and secondary care. If capacity to consent was impaired, a consultee declaration was sought, in line with NHS Health Research Authority guidance.

- Administration of the Modified Telephone Interview for Cognitive Status (TICS-m), two tests of verbal fluency adapted from the Addenbrooke's Cognitive Examination, in which participants were asked to generate words beginning with the same letter and number of animals (Cook, Marsiske et al. 2009, Hsieh, Schubert et al. 2013)

- Medical history, including comorbidities and medication history, general health

- Social history, including educational attainment, current living arrangements, contact with relatives, index of multiple deprivation, health behaviours, hearing, vision, quality of life, dental health, continence, falls, depression, personal and instrumental activities of daily living, care needs,

- Functional status including falls history, continence, hearing, vision, quality of life, dental health, continence, depression, personal and instrumental activities of daily living

- Frailty was quantified as a cumulative index of health deficits (0 to 1), derived using 28 items drawn from the baseline assessment and calculated in line with standard procedures (Searle, Mitnitski et al. 2008), but excluding cognitive items to avoid collinearity with the primary cognitive measure.

Participants were also asked to nominate an informant as an additional source of diagnostic information.

4.4.1.3 Acute admission delirium outcomes and prospective cognitive outcomes

All participants admitted to either UCLH or RFH were automatically flagged to the study team through daily electronic alerts. In addition, the research team screened admissions lists between Monday and Friday identifying participants who had been admitted either electively or as an emergency. Any recruited DELPHIC patients admitted into UCLH or RFH were assessed daily during weekdays by a research team member for the entirety of their inpatient stay. Data recorded included information collected through usual clinical care:

- Demographic: age, sex, education, place of residence, co-resident support
- Clinical: physiological parameters of illness severity (National Early Warning Score, NEWS), illness severity scores (Acute Physiology and Chronic Health Evaluation (APACHE) II (minus arterial blood gas), medications, laboratory findings, presenting complaint
- Cognitive function: Memorial Delirium Assessment Scale (MDAS), Observational Scale of Level of Arousal (OSLA)
- Physical function: Hierarchical Assessment of Balance and Mobility (HABAM) (MacKnight and Rockwood 1995).

For maximal accurate ascertainment during out-of-hours, clinical and delirium-specific data were obtained from medical notes and validated informant delirium measures from ward staff and family trusted advisors (consultees) interviews.

Baseline assessments were repeated two years after initial recruitment. Participants' deaths during the study period were confirmed on local hospital electronic records systems and cross-referenced with the NHS Spine, a statutory register for all deaths in England.

4.4.1.4 Strengths and Limitations

In older people with known baseline function, DELPHIC is the largest dataset of acute illness in terms of number of study participants, inpatient assessment days and admission episodes. It is unique in its cross-setting design, tracking participants across baseline states in the community, primary care and acute illnesses in secondary care. The prospective nature of data collection allows accurate ascertainment of baseline factors without the need for historical recollection, which is frequently suboptimal. Last, the breadth of collected data modalities results in the inclusion of cognitive, functional and social outcomes often not included in epidemiological studies but are critical to consider when studying health trajectories in older people. The majority of measures were collected using validated ordinal scales.

There remain limitations to DELPHIC: first, despite a large number of participants, only a small proportion is hospitalised during the study period. In addition, the study does not have the facility to assess illnesses outside of an inpatient setting, such as acute episodes, which general practitioners in the community alone treat. The same measures of illness severity as in younger populations and intensive care settings have been applied to this study. However, it is unclear whether they are the most optimal scales of illness severity in the older population who may not demonstrate anticipated physiological fluctuations as other age groups. Last, despite contemporaneous linkage of bloods, neuroimaging linkage is currently not available.

4.4.2 UCLH Cognitive Status Dataset

4.4.2.1 Sample

The University College London Hospital (UCLH) dataset comprises of 2951 patients consecutively admitted to the acute Medicine for the Elderly service (seven days/week) at University College London Hospital (UCLH) between March 2015 and March 2017. Recruitment was unselected other than through the indication for the clinical service: an acute general medical problem in a population living with frailty. Eligible patients were those over 65 who had been admitted to UCLH directly by a geriatrician. The admission start date was defined as the date of initial presentation

to the emergency department. The discharge date was defined as the date of the patient leaving UCLH. In the case of patients with multiple admission episodes during the study inclusion period, a unique patient episode was defined as a new admission date recorded more than four weeks after the last blood test investigation. Any re-admission within 4 weeks of their final blood tests was defined in this study as a continuation of the same admission episode, as four weeks was deemed clinically feasible to be within a period of acute illness recovery for older patients. A subset of 1036 unique patient episodes was identified where complete demographics, blood and CT brain data were available. Patients with unknown survival status were excluded. The data were collected as part of a service evaluation project and individual consent was not necessary for these secondary analyses, as determined by the NHS Health Research Authority.

4.4.2.2 Outcomes

Mortality: Inpatient deaths were recorded with the date of death. Data for all patients were censored at December 2018, with an “alive or dead” status recorded for all patients.

4.4.2.3 Exposures

Each patient was reviewed by a consultant geriatrician within 24 hours of hospital admission and clinically classified as having: i) delirium only; ii) dementia only; iii) delirium superimposed on dementia; or iv) no cognitive impairment, from the medical notes and clinical assessment. Dementia was generally diagnosed by medical records or collateral history. Delirium was made as a consultant diagnosis, often but not always using a validated diagnostic tool such as the 4AT. I linked contemporaneous admission information to this clinical dataset, laboratory and imaging investigations, corresponding as closely as possible to the index admission (laboratory results within 48 hours of admission; non-contrast CT head imaging performed within four weeks of admission date). The primary diagnosis of each patient was coded as a chapter header of International Classification of Diseases ICD-10. Each patient’s mortality status and if applicable, date of death, were recorded on 24th December 2018 through the hospital vital statistics database

(Carecast, GE Healthcare). Lengths of stay and vital status (dead/alive) was determined for all patients through clinical care records (Carecast linkage) for a censoring date at 1st December 2018.

4.4.2.4 Study strengths and limitations

The key strength of this dataset is the large-scale linkage of blood and CT images with the oldest-old patients, with a mean age of 85.5 years. Participants were observed for at least 20 months before determining the primary endpoint (mortality) to study medium-term mortality risk. This resulted in roughly equal group sizes of alive or dead individuals at the point of censoring. A limitation is the suboptimal definition of acute cognitive status: using a delirium diagnostic instrument was not compulsory and may have contributed to inconsistency among diagnoses between different consultants. Similarly, for dementia diagnosis, while medical records and collateral histories may be sufficient for many patients, this was not necessarily confirmed with a validation process such as IQCODE or formal corroboration with primary care records. Inherently, both required prior interaction with healthcare services – no dementia diagnosis due to lack of contact with healthcare practitioners did not fully exclude the possibility of clinical dementia. Lastly, the study offers only a partial, cross-sectional snapshot view of relationships between biochemistry and plain CT structural neuroimaging with medium-term mortality risk. Longitudinal changes in biochemistry, neuroimaging and cognitive statuses with later life mortality risk cannot be studied with this dataset.

4.4.2.5 Data Preprocessing

4.4.2.5.1 Blood

Routinely performed tests: full blood count differentials, red cell distribution width, urea, creatinine, glomerular filtration rate, alanine transaminase, alkaline phosphatase, bilirubin, albumin, potassium, C-reactive protein, from the first 48 hours of admission were linked for each admission. Samples from this period encompass the first anticipated venepunctures in emergency care or the acute medical unit. The timeframe can also be reasonably included for a “front door” predictor. Where there were multiple values, I chronologically indexed the first value

and the mean and standard deviation for the admission. Where only one test was performed, first and mean were identical and standard deviation was zero. This procedure yielded a set of 78 variables capturing both static and dynamic changes in each test. Distributions were visually examined, transformed where appropriate, and clipped to enclose values within 99% of the density of the distribution.

Clinical investigations are generally guided by prior, clinically-informed belief. To capture the effect of such "intention to investigate", five levels of investigative intention (II) combined with obtained values were defined for each test: 1) Investigation performed or not performed (1 binary variable); 2) counts of investigation performed over the first 48 hours (1 real-numbered variable); 3) II level 1 and the first test value (2 variables); 4) II level 2 and the mean test value (2 variables); and 5) the first test value, mean and standard deviation (3 variables).

Data were modelled at different II levels to quantify the relative predictive content of the intention to investigate vs the actual test values thereby obtained. It can be interpreted that improving performance using increasing levels of II indicate predictive contributions of the blood values, over and above whether the test has been performed.

4.4.2.5.2 CT Neuroimaging

Non-contrast CT imaging of the head performed within 4 weeks of admission for any indication was linked to each patient episode. Any images of size over 100MB in size or involving a body part other than head were manually excluded. Collaborating groups processed each image within an SPM-based (https://www.fil.ion.ucl.ac.uk/spm/) pipeline that included, in order: rigid-body realignment to Montreal Neurological Institute (MNI) space, resampling to 1mm^3 isotropic resolution, and non-linear unified spatial segmentation and normalisation to MNI space based on a CT- optimised extension of SPM's unified segmentation and normalisation routine (Ashburner and Friston 2005), employing a custom, CT-specific atlas of both intensity and spatial distributions (Blaiotta, Freund et al. 2018) (https://github.com/WCHN/CTseg). This generative model is an extension of the approach implemented in the SPM12 software (Ashburner and Friston 2005), which

allowed for learning an atlas and intensity distributions from populations of brain images (Blaiotta, Freund et al. 2018). The model used in this paper had been trained on a large number of both MR and CT images and is available from https://github.com/WTCN-computational-anatomy-group/diffeo-segment. Segmentations were thresholded at 0.5 for each tissue type.

Scans were resampled at 1.5mm, 3mm, 4.5mm and 6mm isotropic resolutions, extracting all voxels meeting the following criteria: tissue probability >0.5 and voxel-wise probability variance across the cohort >0.01, flattened into a one-dimensional array and horizontally stacked. Voxel intensities were variance clipped at 99%. Any column with only zeros, representing an empty voxel contributing no information input to the final model, were removed. The pipeline output for each patient was two sets of probabilistic tissue segmentation maps of grey matter, white matter, cerebrospinal fluid, skull, and meninges/soft tissue: one native and one non-linearly registered to MNI.

5 Delirium and Population Health Informatics Cohort (DELPHIC) study

Chapter Outline

- Introduction
- Methods
  - Delirium ascertainment
  - Statistical analyses
- Results
  - Study demographics and delirium prevalence
  - Baseline cognition with delirium and arousal measures during illness
  - Baseline cognition and delirium measures with long term cognitive decline and mortality
- Discussion
  - Clinical significance of findings
  - Strengths and limitations
  - Future directions

In this chapter, I will be using methods described in chapter 4.1.1 and 4.1.2 to address evidence gaps in delirium population epidemiology articulated in chapter 3. I will first describe period prevalence of delirium when admitted to a medical ward from a community recruited cohort. Next, I will demonstrate the longitudinal relationships between baseline cognition, acute illness and long-term adverse outcomes in older people using population epidemiology. While dementia and prior cognitive impairment are known risk factors for incident delirium, I will show how baseline cognition is related to delirium severity during acute illness, across the spectrum of premorbid cognitive performance. Next, I will how delirium *per se* is inherently toxic for long-term cognition and mortality, how delirium contributes to both cognitive decline and mortality in a dose-dependent manner, and how this toxic effect is modulated by baseline cognition.

The findings of the association between baseline cognition and delirium phenomenology were published in a manuscript entitled “Extremes of baseline cognitive function determine the severity of delirium: a population study” (Tsui, Yeo et al. 2023). The associations between baseline cognition, delirium and long-term cognitive and mortality outcomes were published in a manuscript entitled “The impact of baseline cognition and delirium on long-term cognitive impairment and mortality: the Delirium and Population Health Informatics Cohort” (Tsui, Searle et al. 2022).

5.1 Introduction

The clinical importance in prompt recognition and appropriate initiation of treatments for delirium is well established. Affecting at least 1 in 4 older admitted patients, presenting as acute changes in arousal, inattention, and global cognitive impairment, delirium has been demonstrated to be associated with adverse outcomes such as mortality, long term cognitive impairment, inpatient falls, delayed discharges, and significant patient/carer distress across a number of clinical settings (Partridge, Martin et al. 2013, Goldberg, Chen et al. 2020, Tieges, Quinn et al. 2021)

Despite this, there remain significant gaps in our current knowledge of the determinants of delirium prevalence risk. Although older age and increased premorbid frailty (Lindroth, Bratzke et al. 2018) have previously been identified, how they interact with concurrent risk factors to result in different cross-sectional delirium severities and cumulative delirium exposures remain unknown. Our limited understanding of the wide variability in the natural history of delirium (Jackson, Wilson et al. 2016) restricts our ability to prognosticate for delirium prevalence, severity and burden, and hence varying degrees of consequent brain injuries.

The influence of baseline cognition on subsequent delirium phenomenology has not been considered comprehensively. Yet, an empirical understanding of this relationship could affect delirium detection and management because the clinical significance of delirium symptoms might have different implications if framed in the context of a known baseline cognitive state.

In addition, the relation of delirium features on adverse outcomes remains poorly understood. An recent expert review of international delirium research emphasised "*particular vulnerabilities predicting the negative long-term outcomes of delirium will be key to identifying at-risk patient groups and to developing targeted therapies*". (Khachaturian, Hayden et al. 2020). Is delirium an epiphenomenon reflective of underlying, likely undiagnosed cognitive impairment or dementia, unmasked during acute illness? Or is delirium inherently toxic per se, resulting in a direct contribution to the adverse effects, independent of baseline cognition, acute illness and hospitalisation? For example, the extent to which cognitive decline observed after hospitalisation is specific to delirium, and not reflective of baseline cognition, is unclear (Mathews, Arnold et al. 2014).

If delirium is an independent contributor to long-term cognitive decline and mortality, it is not known if this manifests in a linear dose-dependent or thresholded manner, and whether other factors such as baseline cognition and frailty may interact to mitigate or worsen these outcomes and trajectories. A better understanding of these relationships has significant clinical implications because it would highlight those at greatest risk of delirium and its adverse effects. It would identify those who would benefit most from the intensive delirium prevention and post-delirium follow-up.

Prior studies linking baseline cognitive function to delirium have used the methodological advantage of prospective follow-up in elective surgical populations (Saczynski, Marcantonio et al. 2012, Wu, Shi et al. 2015, Lindroth, Bratzke et al. 2019). However, most delirium in secondary care presents in unselected unscheduled admissions in whom there is a much greater range of pre-existing cognitive impairment and frailty (Ahmed, Leurent et al. 2014). To encompass the whole spectrum of all delirium acute hospital presentations, we needed to characterise cognitive function in a stable community sample. Then, at each acute hospitalisation objectively: (1) assess the point-prevalence, severity and duration of delirium; (2) quantify these in relation to baseline cognitive function; (3) estimate these associations with to long-term cognitive and mortality outcomes. We hypothesised first that baseline cognition would contribute to differing delirium risk, severity and duration; and second, that baseline cognition would interact with the

cumulative delirium exposure, resulting in different long-term cognitive and mortality outcomes.

5.2 Methods

5.2.1 Ascertainment of delirium

Case ascertainment of delirium was diagnosed using the Diagnostic and Statistical Manual of Mental Disorders (DSM) IV criteria. The wide use of DSM IV criteria within research literature allows easy comparison with other studies. Delirium was determined to be present if individuals met criteria A (*disturbance of consciousness*), B (*change in cognition and/or perception*) and C (*acute and fluctuates*). By virtue of their inpatient admission, all participants were deemed to fulfil Criterion D (*physiological consequence of a general medical condition*). Operationalisation of DSM-IV criteria and their corresponding interview questions are demonstrated in table 5.3.

5.2.2 Statistical analyses

5.2.2.1 Definition of outcome measures

5.2.2.1.1 Delirium

Delirium ascertainment was performed daily for each admitted participant. Delirium prevalence was defined as:

(i) daily prevalence (delirium on each specific day of admission);

(ii) admission prevalence (delirium at any point during admission)

(iii) study period prevalence (delirium at any point during the participant's inclusion in the DELPHIC study).

(iv) *Point prevalence*: proportion of total study participant inpatient days positive for all DSM-IV components

In addition to prevalence, delirium was further operationalised as:

a) *Severity*: *Daily scores of Memorial Delirium Assessment Scale (*MDAS) assessing 10 domains of delirium symptoms (each scored out of 3) to give a 30-point measure of daily delirium severity

b) *Duration*: total number of delirium days experienced over the study period (across multiple admissions where relevant).

c) *Burden*: Total delirium burden was quantified by combining a measure of duration and severity through summing daily MDAS scores (expressed as point*days).

To allow for more robust analyses and easier interpretation, a categorical outcome of delirium burden was further constructed. This anticipates a long right tail in the distribution of delirium burden, contributed to by patients with long admission durations. Admitted individuals with scores above the median were classed as 'high delirium burden'; those with scores below median classed as 'low delirium burden'. All other participants were classified as 'no delirium burden'.

5.2.2.1.2 Cognitive function and death

a) *Cognitive function*:

A composite score was constructed using the sum of TICS-m (total 53 points) and verbal fluency tasks (two tasks scoring up to 7, total 14 points). This was standardised and distribution noted to be Gaussian with mean of 0; +2.0 SD and -2.0 SD including more than 95% of values. The same composite score was used for both baseline cognition (as an exposure) and follow-up cognition (as an outcome).

b) *Mortality*

Any deaths occurring before cognitive assessment follow up, either as an inpatient in hospital or in the community.

5.2.2.1.3 Missing data

In-hospital assessments missing due to falling on a weekend or public holiday (missing at random) were forward-filled (Friday carried to Saturday) and backward-

filled (Sunday carried from Monday) in 24-hour intervals for up to 4 days. For backward filling, this approach has the advantage of automatically carrying over information from the next working day's chart review. Otherwise, data were assumed to be missing at random.

5.2.2.2 Models:

5.2.2.2.1 Baseline cognition and subsequent delirium and arousal:

Mixed effects regression models estimated associations between baseline cognition and subsequent delirium measures, using logistic regression for delirium prevalence and linear regression for delirium severity. Models were clustered by participant study ID. Negative binomial regression was used to model delirium duration. Days in delirium was operationalised as count data over the number of days spent in hospital over the study period. All analyses were adjusted by standardised age, sex, baseline cognition, educational attainment, frailty index and mean daily NEWS. Standardisation for continuous, normally distributed data was calculated as (score-mean)/standard deviation.

To investigate non-linear relationships between cognition and delirium severity and abnormal arousal, we fitted restricted cubic splines with three knots. Default knot positioning from the Stata "mkspline" function was used, which operationalises Harrell's recommended percentiles with the additional restriction that the smallest knot may not be less than the fifth-smallest value of baseline cognition and the largest knot not be greater than the fifth-largest value of baseline cognition.

Sensitivity analysis:

It was noted that specific outcome items, such as MDAS items 2 (disorientation) and 3 (short-term memory impairment), may score more highly due to premorbid cognitive impairment, regardless of delirium exposure. In order to further highlight delirium specific contributions to adverse outcomes, over and above possible pre-existing cognitive impairment or dementia, sensitivity analyses using principal models were also thus performed using cognitive composite scores with these two items removed (modified outcome score /24 points).

5.2.2.2.2 Delirium and long-term cognitive impairment or death:

Linear regression estimated change in cognition at follow up. Mortality risk was estimated using Cox regression to model proportional hazards of death. For both models, incident delirium (yes/no) and then a categorised delirium burden variable (none/low/high) were used as the independent variable, with age, sex, baseline cognition, educational attainment, frailty index as standardised cofounders. Models performed at an admission episode level utilised mean NEWS as the summary measure of illness severity. Lastly, in order to delineate whether delirium measures per se, incidence or burden, resulted in any outcome changes or whether they were related to hospitalisation more broadly instead, sensitivity analyses were performed using length of hospital admission as an independent variable.

5.3 Results

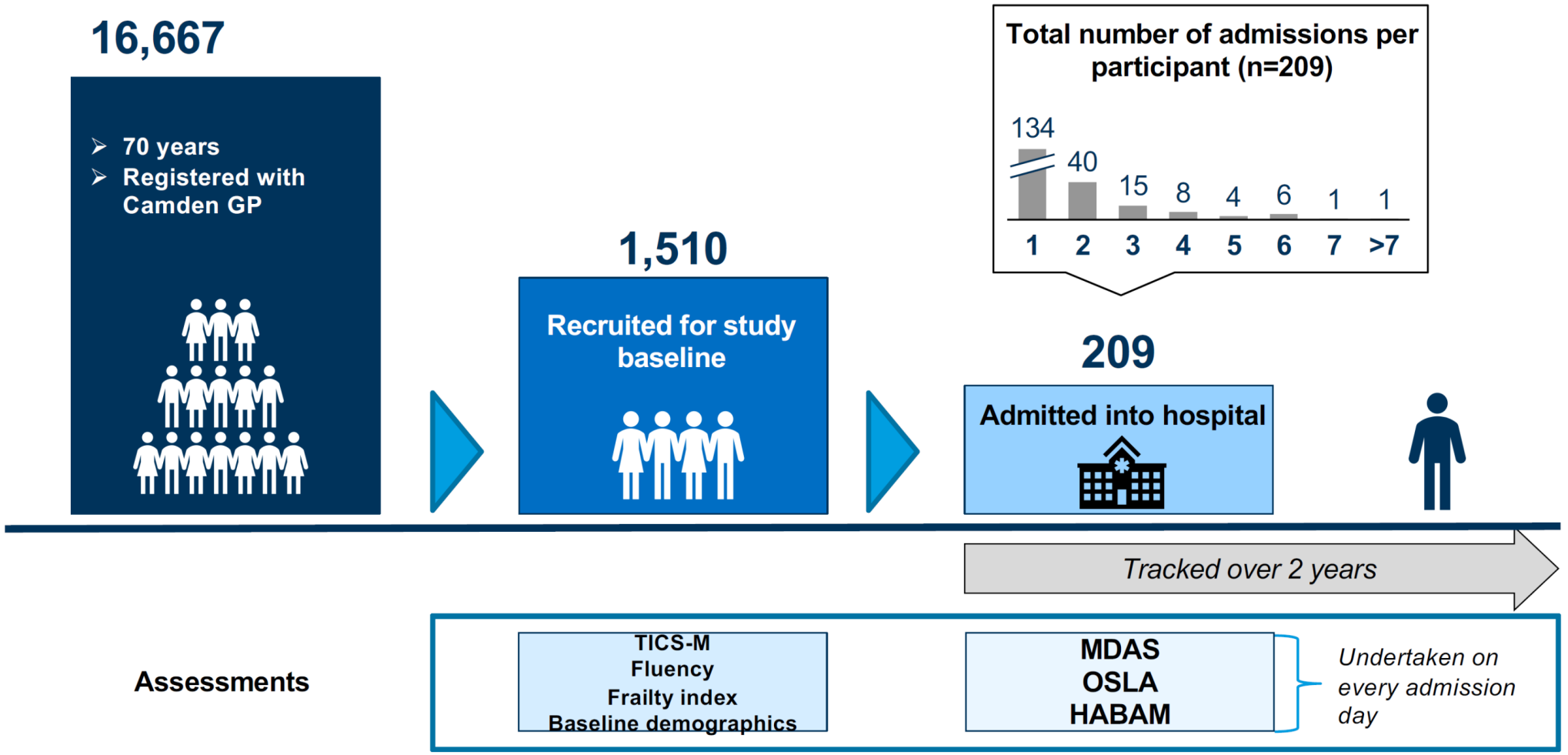
5.3.1 Demographics

Of 1510 participants recruited between March 2017 and October 2018, median age was 77 (interquartile range 73 to 82), and 57% were women (Table 5.1)

Home assessments were undertaken for 32 participants who were not able to use a telephone. Compared with the demographics of Camden over the age of 70, the sample was well matched by age. The absolute differences were small except for ethnic representation, in which DELPHIC contained a greater proportion of people of White ethnicity (table 5.1, fig 5.1) and Index of Multiple Deprivation, where participants in the DELPHIC scored lower than the median for the Borough of Camden residents (table 5.2).

Over the study period, 209 participants (14%) were hospitalised across 371 episodes, representing 1999 person-days of assessment.

Figure 5.1: Cohort structure showing sample and schedule of assessments



1510 participants were recruited from 16,667 eligible patients registered with a GP in Camden. Cognition was assessed with a combined TICS-M and fluency measure at baseline and follow up. Measures of delirium severity (MDAS – Memorial Delirium Assessment Scale), arousal (OSLA – Observation Scale of Level of Arousal) and function (HABAM – Hierarchical Assessment of Balance and Mobility) were assessed daily during inpatient admission stays

Table 5.1: Characteristics of the cohort in relation to hospitalisation and delirium status

	Whole Cohort	Delirium		Attrition		Died		Followed up	
	n = 1510	n = 115	p	n=199	p	n=93	p	n=1218	p
Age	78 (6·2)	82 (6·6)	<0.0001	80 (6·8)	<0·0001	83 (5·8)	<0·0001	77 (5·6)	0·57
Women	57%	55%	0.581	60%	0·36	46%	0·03	58%	<0·0001
Education					<0·0001		<0·0001		<0·0001
Degree level	65%	40%		54%		40%		68%	
Up to secondary (12y schooling)	21%	30%							
Up to primary (6y schooling)	14%	30%	<0.0001	21%		33%		12%	
White ethnicity	94%	89%	0.005	91%	0·12	95%	0·89	95%	0·26
Frailty Index (SD)	0·15 (0·13)	0·30 (0·17)	<0.0001	0·2 (0·17)	<0·0001	0·30 (0·1)	<0·0001	0·13 (0·1)	<0·0001
TICS-m (total, SD)	38·8 (5·9)	33·8 (8·7)	<0.0001	36 (7·5)	<0·0001	34 (4·9)	<0·0001	40 (4·9)	<0·0001
Fluency (words, SD)	15·6 (6·2)	11·6 (6·8)	<0.0001	14 (6·1)	<0·0001	11 (6·8)	<0·0001	16 (6·0)	<0·0001
Fluency (animals, SD)	19·0 (7·0)	13·3 (7·4)	<0.0001	17 (7·5)	<0·0001	13 (6·5)	<0·0001	20 (6·5)	<0·0001
Self-rated health (poor/very poor)	18%	49%	<0.0001	24%	<0·0001	46%	<0·0001	14%	<0·0001
Past medical history									<0·0001
Myocardial infarction	21%	37%	<0.0001	22%	0·59	36%	<0·0001	<0·0001	<0·01
Diabetes mellitus	12%	19%	0.019	15%	0·17	25%	<0·0001	<0·0001	<0·01
Hypertension	50%	61%	0.01	48%	0·61	67%	<0·0001	31%	0·11
Stroke	9%	16%	0.007	11%	0·34	20%	<0·0001	8%	<0·0001
Cancer	24%	25%	0.644	19%	0·08	29%	0·19	24%	0·49
COPD	14%	28%	<0.0001	17%	0·13	<0·0001	<0·01	12%	<0·0001
Any impaired PADL	9%	31%	<0.0001	18%	<0·0001	36%	<0·0001	6%	<0·0001
Any impaired IADL	73%	90%	<0.0001	77%	<0·0001	86%	<0·0001	73%	<0·0001

P values refer to the following comparisons: *Delirium* cf whole cohort; *Attrition* cf followed-up; *Died* cf followed up; *Followed up* cf whole cohort.
TICS-m Modified Telephone Interview for Cognitive Status. PADL Personal activities of daily living: grooming; toileting; dressing; bathing; transfer; stairs; IADL Instrumental activities of daily living: shopping; washing up; making hot drink; feeding; walking outside

For participants presenting to hospital, the commonest presenting symptoms were general malaise and fever (15%), respiratory (dyspnoea, cough, 9%) and neurological (confusion, 9%) complaints. Hospitalised individuals had lower baseline TICS-m cognitive scores (mean 35.5 versus 38.8 points, p<0.01), and more frailty (frailty index 0.25 versus 0.15, p<0.01) than those not hospitalised. Individuals admitted once accounted for 114 (55%) hospital episodes, and the rest were admitted multiple times (median number recurrent admissions 2, IQR 2 to 4).

Table 5.2: Demographic characteristics of the DELPHIC sample in relation to the London Borough of Camden

	Camden	**DELPHIC**
Age (median, IQR)	77 (73 to 83)	77 (73 to 82)
Female	58%	57%
Index of multiple deprivation (median, IQR)	18.9 (11.3 to 28.4)	14.6 (9.1 to 22)
Ethnicity (% white)	84%	94%

Although follow up was initially planned for two years after initial recruitment, the Covid-19 pandemic resulted in a delay to cognitive follow ups, particularly in those requiring a home visit. As a result, the final median follow-up period was 3.5 years, totally 5059 person.years to July 2021. By the end of the study, 1218 participants (81%) had repeat cognitive assessments, 113 (8%) had died, and 180 were lost (or withdrew) to follow up (12%). There were significant differences between participants who completed follow up compared to those who were lost: those with lower baseline cognition were more likely to be lost to follow up (OR -0.82 per SD baseline cognitive score; 95%CI -1.28 to -0.38; p<0.01), as were male participants, those were lower educational attainment, frailer participants at baseline, have poorer general health and more dependent on activities of daily living.

5.3.2 Delirium prevalence

On any given day (point prevalence), 29% of hospitalised participants fulfilled DSM-IV criteria for delirium. At any assessment, participants met DSM-IV criteria A, B and C 69%, 68%, and 41% of the time (table 8.3). Measures contributing to Criterion A included abnormal OSLA scores (31%) and inability to perform *months of the year* backward (13%). Components of Criterion B included short-term memory impairment in 31% of cases and perceptual disturbance in 13%. There was evidence of fluctuation (Criterion C) in OSLA or MDAS scores (differing from the previous assessment by ≥1 SD) 5% of the time. New severe sleep-wake cycle disturbance was present in 20%.

Median MDAS score was 7/30 points (IQR 3-12). Delirium burden (cumulative MDAS scores) had a median 26 point*days (IQR 20-197), equivalent to around three days of moderate delirium. Over the course of an admission, delirium was ascertained at any point within the episode in 45% of admitted participants (prevalent delirium at admission in 35%, incident delirium developing after admission in 10%). The average number of days with delirium (consecutively positive assessments) was 3.9 days. In 61.6% of admission, the participant presented with delirium on the day of admission. In 44.5% of admissions, the participant remained delirious according to DSM-IV criteria on the final day of assessment, assumed to be the day of discharge.

Table 5.3: Point prevalence of delirium features in hospitalised sample contributing to DSM-IV case ascertainment.

Criterion A		**Criterion B**		**Criterion C**	
69%		**68%**		**41%**	
Item 1 ≥2: reduced level of consciousness	33%	Item 2 ≥2: disorientation (time/place questions 5/10 errors)	32%	Item 10 ≥3: sleep-wake cycle disturbance	17%
Item 4 ≥2: impaired digit span (5 forward or 3 backward errors)	10%	Item 3 ≥2: short-term memory impairment (≥2 errors on 3-item delayed recall)	31%	Observed fluctuations in arousal	6%
Item 5 ≥2: inattention	30%	Item 6 ≥2: disorganized thinking	15%	Observed motor fluctuations	5%
Inattention during interview	4%	Item 7 ≥2: perceptual disturbance	13%	Informant report of fluctuations	22%
Dozes off during interview	1%	Item 8 ≥2: delusions	25%	Item or OSLA score different from previous assessment by ≥1 SD	5%
Distracted by environmental stimuli	3%	Informant report *more confused*	7%		
OSLA total ≥2	31%	*Odd thoughts* described on direct questioning	2%		
MOTYB >5 mistakes	13%	Hallucinations described on direct questioning	3%		
Serial 7 score lower than baseline	16%	*Strange things* described on direct questioning	1%		
		3 sentences to complete (3 choice answer) (any error)	8%		
		2 sentences to complete (free choice answer) (either error)	7%		
		2-stage sequencing command (either error)	7%		

Each MDAS item is rated 0, 1, 2 or 3. Criterion present if one or more symptom/sign positive

Note MDAS item 9 (decreased or increased psychomotor activity) is not used in the case definition.

OSLA Observational Scale for Level of Arousal; MOTYB months of the year backward; Informants: health care staff and/or family/carers.

5.3.3 Baseline association with delirium measures

5.3.3.1 Delirium incidence

In univariate models, poorer baseline cognition and increased frailty were both associated with increased risk of incident inpatient delirium if admitted during the study period. For each standard deviation of increased baseline cognition, the participant was 44% less likely to develop incident delirium (OR 0.56 95% CI 0.42 to 0.74, p value <0.01). Correspondingly, for each standard deviation of increased baseline cognition, the participant was 65% more likely to develop incident delirium (OR 1.65, 95% CI 1.22 to 2.24, p value <0.01).

After adjustment for age, sex, educational attainment, NEWS and frailty, the association between baseline cognition and odds of incident delirium slightly attenuated (OR 0.63 per SD, 95%CI 0.45-0.89, p=0.01) (fig 8.2, table 8.4). However, the association between frailty and incident delirium attenuated on full adjustment (OR 1.27 per SD, 95%CI 0.89-1.81, p=0.18).

Table 5.4: Associations between baseline cognition and delirium incidence (DSM IV status)

		Univariate Analyses				Multivariate Analyses			
n = 1999, 209 clusters		**OR**	**95% CI**		**P value**	**OR**	**95% CI**		**P value**
Baseline Cognition*		0.56	0.42	0.74	<0.0001	0.63	0.45	0.89	0.01
Education									
	Degree level								
	Up to secondary (12y schooling)	1.32	0.59	2.92		1.49	0.67	3.28	
	Up to primary (6y schooling)	0.69	0.34	1.43	0.23	1.09	0.52	2.26	0.59
Age		1.04	0.99	1.08	0.14	0.98	0.7	1.38	0.9
Female		0.94	0.52	1.71	0.84	0.9	0.5	1.61	0.72
FI (minus cog)		1.65	1.22	2.24	<0.0001	1.27	0.89	1.8	0.18
NEWS		1.03	0.95	1.12	0.5	0.89	0.66	1.22	0.48

Mixed-effects logistic regression accounts for repeated measures per individual. Baseline cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score. Univariable analyses are individual models per row; multivariable analyses show coefficients mutually adjusted for all other factors.

5.3.3.2 Delirium severity

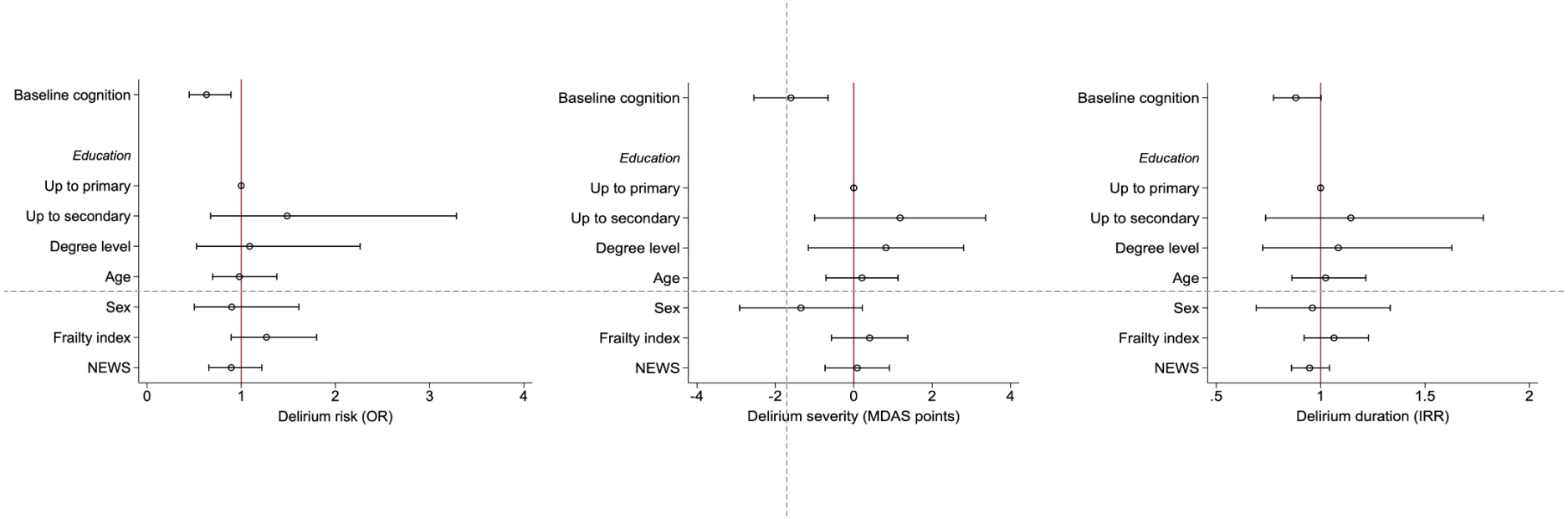
In univariate analyses, poorer baseline cognition, older age and greater frailty were all associated with more severe delirium. Each standard deviation decrease in baseline cognition was associated with a 1.8 point increase in MDAS score (coefficient -1.8, 95% CI -2.57 to -1.03, p <0.0001). However, educational attainment was not associated with delirium severity. On full adjustment, worse baseline cognition was associated with more severe delirium (-1.6 MDAS point per SD baseline cognitive score, 95%CI -2.55 to -0.66, p<0.0001) (fig 5.2, table 5.5**)**.

Table 5.5: Associations between baseline cognition and delirium severity (Memorial Delirium Assessment Scale score)

	Univariate Analyses				Multivariate Analyses			
n = 1999, 209 clusters	**Coefficient**	**95% CI**		**P value**	**Coefficient**	**95% CI**		**P value**
Baseline Cognition*	-1.80	-2.57	-1.03	<0.0001	-1.60	-2.55	-0.66	<0.0001
Education								
Degree level								
Up to secondary (12y schooling)	0.91	-1.33	3.15		1.19	-0.95	3.33	
Up to primary (6y schooling)	-0.58	-2.54	1.38	0.35	0.82	-1.16	2.80	0.53
Age	0.13	0.00	0.26	0.05	0.03	-0.10	0.17	0.66
Female	-1.20	-2.84	0.43	0.15	-1.35	-2.91	0.22	0.09
FI (minus cog)	1.35	0.52	2.18	<0.0001	0.41	-0.57	1.38	0.41
NEWS	0.08	-0.10	0.26	0.37	0.08	-0.10	0.25	0.39

Mixed-effects linear regression accounts for repeated measures per individual. Baseline cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score. Univariable analyses are individual models per row; multivariable analyses show coefficients mutually adjusted for all other factors.

Figure 5.2. Relationship between baseline factors and delirium risk, severity and duration



Mixed effects multivariable-adjusted estimates showing better baseline cognition is associated with lower probability of delirium and less severe delirium; negative binomial regression multivariate-adjusted estimates showing better baseline cognition is associated shorter duration of delirium.

Left panel: delirium risk (odds ratio); *Middle panel*: delirium severity (MDAS severity score, out of 30); *Right panel*: Delirium duration (incidence rate ratio). Data presented in Supplementary Tables 3-5. Baseline cognition defined by the modified Telephone Interview for Cognitive Status, augmented by two verbal fluency tasks. NEWS: National Early Warning Score. Baseline cognition, frailty index and NEWS standardised to show comparable effect sizes

There was a non-linear relationship between baseline cognition and delirium severity. MDAS scores were higher when baseline cognition was both low and high. The negative relationship between baseline cognition and delirium severity for the first spline and positive relationship with the second spline was reflected in MDAS score of 15 (95% CI 12 to 17) points at *z*-score = -2 and MDAS of score 11, 95% CI 7 to 15 at *z*-score = +2) ((table 5.6, fig 5.3). The lowest MDAS severity scores were seen in those at the midpoint of baseline cognition (*z*-score = 0).

Sensitivity analyses using only non-cognitive items from the MDAS showed similar a similar bimodal distribution of scores (table 5.7 and 5.4).

5.3.3.3 Baseline cognition and abnormal arousal

The relationship between baseline cognition and abnormal arousal followed a comparable pattern. At the extremes of baseline cognition, OSLA scores were higher (OSLA 4, 95% CI 3 to 6 points at *z*-score = -2; OSLA 4, 95% CI 2 to 7 at *z*-score = +2) (Figure 5.3). Again, the lowest OSLA scores were recorded in those with baseline cognition *z*-scores of 0.

Table 5.6: Non-linear relationship between delirium severity (MDAS score) or arousal abnormality (OSLA score) and baseline cognition, as demonstrated by multivariate non-linear polynomial regression using restricted cubic spline with 3 knots

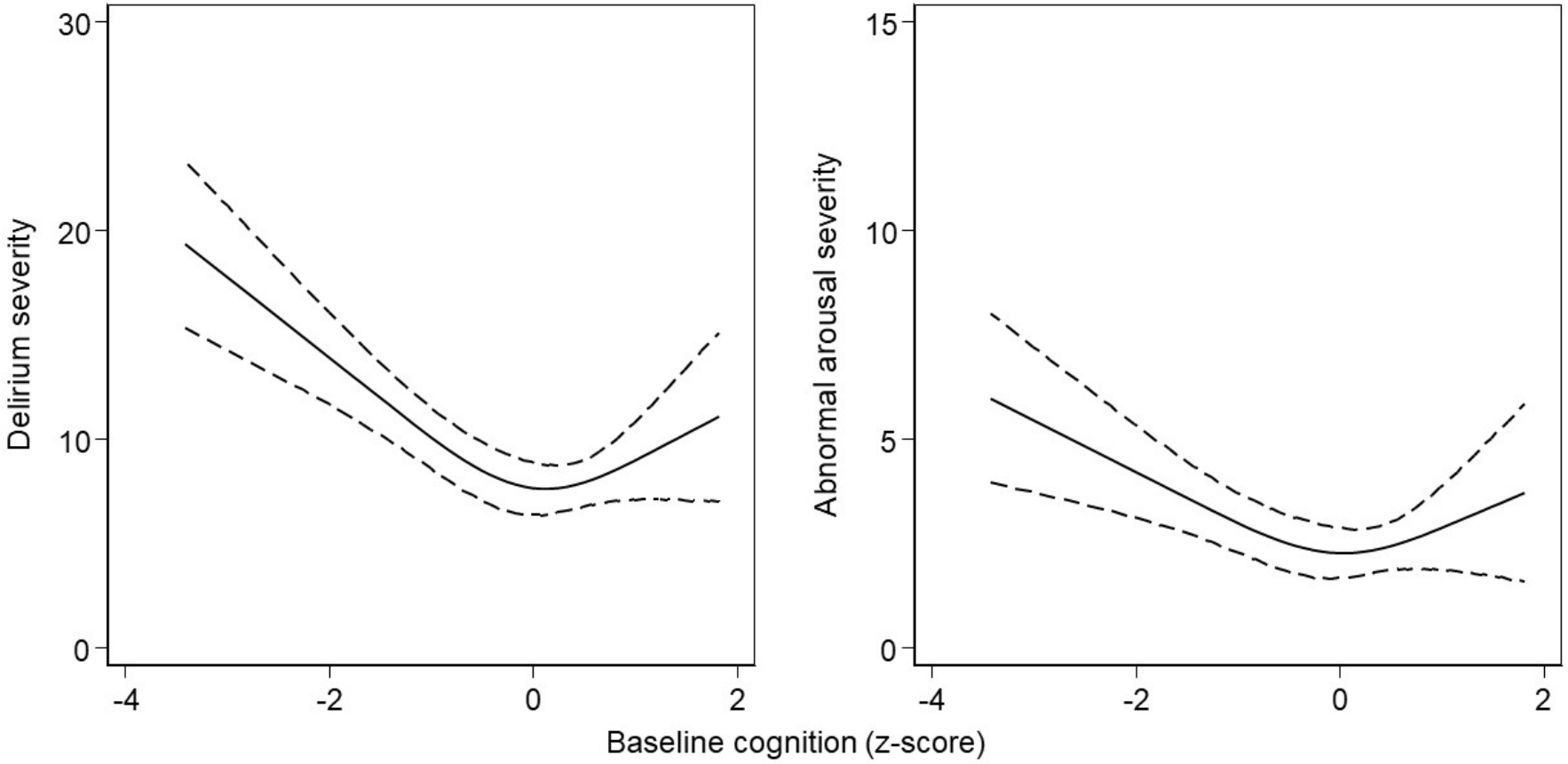
	MDAS				OSLA			
	β	**95% CI**		**P**	**β**	**95% CI**		**P**
Cognition: first spline	-3.46	-4.59	-2.33	<0.01	-1.07	-1.77	-0.36	<0.01
Cognition: second spline	3.14	1.18	5.09	<0.01	0.97	-0.04	1.98	0.06
Age (per SD)	0.35	-0.87	1.58	0.57	0.20	-0.42	0.83	0.52
Sex (women vs men)	-1.10	-3.05	0.86	0.27	-0.37	-1.42	0.68	0.49
Educational attainment*				0.68*				0.82*
Up to primary (6 years)	Ref				Ref			
Up to secondary (12 years)	0.38	-2.15	2.91		0.19	-0.95	1.33	
Degree or higher	0.99	-1.28	3.26		0.38	-0.84	1.59	
Frailty index (per SD)	0.57	-0.55	1.68	0.32	0.10	-0.47	0.67	0.73
NEWS (per SD)	-0.22	-0.77	0.32	0.42	0.12	-0.18	0.42	0.43
Time to first assessment (months)	0.18	0.03	0.32	0.02	0.09	0.02	0.16	0.01

NEWS National Early Warning Score

* Education p values for trend

** estimates also adjusted by age, sex, educational attainment, frailty, NEWS and time to first assessment

Figure 5.3: Bimodal relationship between delirium severity and abnormal arousal with baseline cognition



Trajectories of change in delirium severity scores showing more severe delirium throughout admission for those with lower baseline cognition. Plots are of splines fitted to median value in each tertile of baseline cognition, defined by the modified Telephone Interview for Cognitive Status and augmented by two verbal fluency tasks. MDAS = Memorial Delirium Assessment Scale.

Table 5.7: Non-linear relationship between sub-cortical delirium severity (MDAS score without items 2 and 3) and baseline cognition, as demonstrated by multivariate non-linear polynomial regression using restricted cubic spline with 3 knots

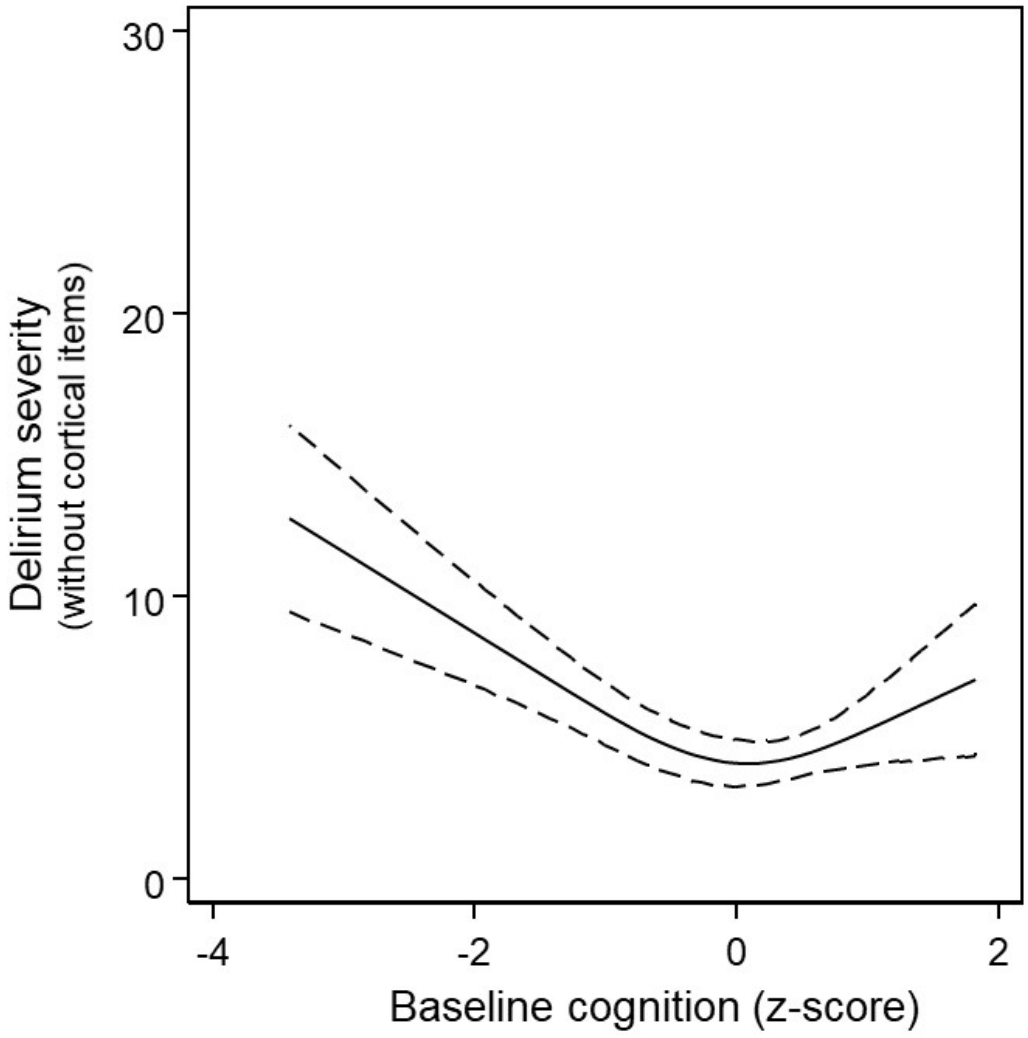
	MDAS (without cortical items)			
	β	95% CI		P
Cognition: first spline	-2.47	-3.39	-1.55	<0.01
Cognition: second spline	2.49	1.10	3.89	<0.01
Age (per SD)	0.32	-0.54	1.19	0.46
Sex (women vs men)	-0.78	-2.12	0.56	0.25
Educational attainment*				0.68*
Up to primary (6 years)	Ref			
Up to secondary (12 years)	0.38	-1.42	2.18	
Degree or higher	0.59	-1.00	2.19	0.76
Frailty index (per SD)	0.24	-0.57	1.06	0.56
NEWS (per SD)	-0.13	-0.53	0.27	0.53
Time to first assessment (months)	0.00	0.00	0.01	0.10

NEWS National Early Warning Score

* Education p values for trend

** estimates also adjusted by age, sex, educational attainment, frailty, NEWS and time to first assessment

Figure 5.4: Bimodal relationship between delirium severity as demonstrated without cortical items with baseline cognition



Trajectories of change in delirium severity scores without MDAS items 2 and 3, showing more severe delirium throughout admission for those with lower baseline cognition, similar to figure Y with full inclusion of MDAS items. Plots are of splines fitted to median value in each tertile of baseline cognition, defined by the modified Telephone Interview for Cognitive Status and augmented by two verbal fluency tasks. MDAS = Memorial Delirium Assessment Scale.

5.3.3.4 Delirium duration

In univariate analyses, poorer baseline cognition and increased frailty were associated with longer duration of delirium. Each standard deviation decrease in baseline cognition was associated with 15% increased rate of delirium incidence. Similarly, each standard deviation increase in frailty index was associated with 17% increased rate of delirium incidence. Educational attainment was not associated with delirium duration. On full adjustment, the associated between baseline cognition and delirium duration was maintained (IRR 0.88, 85% CI 0.77 to 1.00 p = 0.05), while the association for frailty is attenuated (table 5.8). Clinically, this translates to a patient with one standard deviation better baseline cognition experiencing one fewer day of delirium per week of illness than patients of comparable frailty and illness severity.

Table 5.8: Univariate and multivariate associations between baseline cognition and delirium duration

	Univariate Analyses				Multivariate Analyses			
n = 209	**IRR**	**95% CI**		**P value**	**IRR**	**95% CI**		**P value**
Baseline Cognition*	0.85	0.77	0.94	<0.0001	0.88	0.77	1.00	0.05
Education								
Degree level								
Up to secondary (12y schooling)	1.14	0.73	1.76		1.14	0.74	1.78	
Up to primary (6y schooling)	0.89	0.60	1.32	0.50	1.09	0.72	1.63	0.83
Age*	1.02	0.99	1.04	0.17	1.00	0.98	1.03	0.79
Female	1.00	0.72	1.39	0.99	0.96	0.69	1.33	0.81
FI (minus cog)*	1.17	1.04	1.32	0.01	1.06	0.92	1.23	0.40
NEWS	0.95	0.86	1.04	0.25	0.95	0.86	1.04	0.27

Mixed-effects linear regression accounts for repeated measures per individual. IRR: incidence rate ratio. Baseline cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score. Univariable analyses are individual models per row; multivariable analyses show coefficients mutually adjusted for all other factors.

5.3.3.5 Delirium burden

Delirium burden was subsequently calculated as the cumulative sum of daily MDAS scores while the participant was an inpatient within the duration of the study.

In univariate analyses, better baseline cognition, younger age, reduced frailty and reduce illness severity were associated with lower delirium burden (table 5.9). Specifically in relation to baseline cognition, for every standard deviation increase in baseline cognition, this was associated with a 23% (OR 0.77, 95% CI 0.62 to 0.94) and 60% (OR 0.4, 95% CI 0.34 to 0.48) reduction in low and high delirium burden respectively (p value for trend <0.0001). On full adjustment, baseline cognition was not associated with the risk of low delirium burden but one standard deviation was associated with 38% reduced risk of high delirium burden (OR 0.62, 95% CI 0.47 to 0.81, p value <0.0001 for trend). Associations for frailty and illness severity as operationalised by mean NEWS score were maintained on multivariate analyses. However, while significant associations between increasing age and educational attainment with greater delirium burden were demonstrated in univariate analyses, this was not maintained in multivariate analyses.

Table 5.9: Univariate and multivariate associations between baseline cognition and delirium burden category

		Univariate Analyses				Multivariate Analyses			
n = 209	**Delirium Burden**	**Coefficient**	**95% CI**		**P value**	**Coefficient**	**95% CI**		**P value**
Baseline Cognition*	None	Ref				Ref			
	Low	0.77	0.62	0.94		1.00	0.72	1.38	
	High	0.40	0.34	0.48	<0.0001	0.62	0.47	0.82	<0.0001
Education**									
	None	Ref				Ref			
Degree level	Low	Ref							
Up to secondary (12y schooling)		0.89	0.48	1.65		0.92	0.37	2.26	
Up to primary (6y schooling)		0.58	0.33	1.00		0.95	0.43	2.12	
Degree level	High	Ref							
Up to secondary (12y schooling)		0.48	0.27	0.83		0.63	0.25	1.60	
Up to primary (6y schooling)		0.24	0.15	0.38	<0.0001	0.93	0.42	2.05	0.86
Age	None	Ref				Ref			
	Low	1.28	1.05	1.56		1.01	0.96	1.06	
	High	2.09	1.74	2.51	<0.0001	1.05	1.00	1.10	0.16
Female	None	Ref				Ref			
	Low	1.07	0.71	1.62		1.13	0.64	1.99	
	High	0.73	0.49	1.10	<0.0001	0.73	0.39	1.37	0.38
FI (minus cog)*	None	Ref				Ref			
	Low	1.50	1.25	1.81		1.13	0.83	1.53	
	High	2.52	2.14	2.97	<0.0001	1.55	1.18	2.04	0.01
NEWS (mean)*	None	Ref				Ref			
	Low	95.65	40.39	226.52		72.24	31.33	166.53	
	High	100.82	42.53	238.96	<0.0001	78.71	33.98	182.34	<0.0001

Mixed-effects linear regression accounts for repeated measures per individual. Baseline cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; Delirium burden defined as none, low (below median = 26) and high (above median >26); FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score. Univariable analyses are individual models per row; multivariable analyses show coefficients mutually adjusted for all other factors.

*Baseline cognition, FI and mean NEWS standardised per SD, age represented per year

** Education p value for trend

5.3.4 Delirium measures and long-term adverse outcomes

5.3.4.1 Long-term cognitive decline

Delirium diagnosed at any point was associated with worse long-term cognitive decline. Associations between delirium prevalence and long-term cognition were maintained from univariate analyses (coefficient -0.83, 95% CI -1.1 to -0.56, p<0.01) to multivariate analyses (coefficient -0.34, 95% CI -0.62 to -0.06, p = 0.02). Associations with worse long-term cognition were also demonstrated in multivariate analyses with poorer baseline cognition (coefficient 0.59 per SD baseline cognition, 95% CI 0.52 to 0.66, p<0.0001), older age (coefficient -0.02, 95% CI -0.03 to -0.01, p<0.0001), and educational attainment (degree level education coefficient 0.31 95% CI 0.16 to 0.47 p<0.0001) (Table 5.10). No associations were demonstrated between long term cognitive decline with sex, frailty, illness severity during study as operationalised by mean NEWS on full adjustment.

Beyond the general association between delirium incidence and follow-up cognition, there were different effect sizes according to baseline cognition (interaction term p=0.05, table 8.10). While participants with low baseline cognition had similar scores at follow up (z score -1.25, -1.26 for delirium absent and present respectively), the degree of decline at follow up was marked for those at medium baseline cognition (-0.08 and -0.42 for delirium absent and present respectively) and high baseline cognition (1.10 and 0.43 for delirium absent and present respectively).

Table 5.10: Associations between delirium prevalence and follow up cognition

	Univariate Analyses				Multivariate Analyses			
n = 1218	**Coefficient**	**95% CI**		**P value**	**Coefficient**	**95% CI**		**P value**
Delirium Y/N	-0.83	-1.10	-0.56	<0.0001	-0.34	-0.62	-0.06	0.02
Baseline Cognition*	0.71	0.65	0.76	<0.0001	0.59	0.52	0.66	<0.0001
Delirium Y/N - Baseline Cog interaction					-0.16	-0.33	0.00	0.05
Education								
Up to primary (6y schooling)								
Up to secondary (12y schooling)	0.66	0.47	0.86		0.14	-0.03	0.31	
Degree level	1.06	0.89	1.22	<0.0001	0.31	0.16	0.47	<0.0001
Age	-0.06	-0.07	-0.05	<0.0001	-0.02	-0.03	-0.01	<0.0001
Female	0.10	-0.02	0.21	0.09	0.06	-0.03	0.15	0.19
FI (minus cog)*	-0.46	-0.52	-0.39	<0.0001	-0.06	-0.12	0.01	0.09
NEWS (mean)*	-0.09	-0.16	-0.03	0.01	0.02	-0.04	0.08	0.55

Univariate and multivariate linear regression analyses. Baseline and follow up cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score, calculated as mean over study admission days. Multivariable analyses show coefficients mutually adjusted for all other factors.

*Baseline cognition, FI and mean NEWS standardised per SD, age represented per year

** Education p value for trend

Table 5.11: Associations between delirium burden and follow up cognition

n = 1218		Univariate Analyses				Multivariate Analyses			
		Coefficient	**95%CI**		**P value**	**Coefficient**	**95%CI**		**P value**
Delirium Burden	None	Ref				Ref			
	Low	-0.13	-0.36	0.10		0.05	-0.18	0.28	
	High	-1.10	-1.42	-0.78	<0.0001	-0.60	-0.94	-0.26	<0.0001
Baseline Cognition		0.71	0.65	0.76	<0.0001	0.63	0.56	0.70	<0.0001
Delirium Burden - Baseline Cog interaction	None					Ref			
	Low					0.07	-0.12	0.26	
	High					-0.39	-0.60	-0.18	<0.0001
Age		-0.06	-0.07	-0.05	<0.0001	-0.02	-0.03	-0.01	<0.0001
Sex		0.10	-0.02	0.21	0.09	0.03	-0.06	0.13	0.45
FI		-0.46	-0.52	-0.39	<0.0001	-0.08	-0.14	-0.01	0.02
NEWS		-0.09	-0.16	-0.03	0.01	0.00	-0.07	0.07	0.99

Linear regression analyses. Baseline and follow up cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score, calculated as mean over study admission days. Delirium burden derived by thresholding cumulative MDAS scores of participants who experienced delirium at median. Multivariable analyses show coefficients mutually adjusted for all other factors. Baseline cognition, FI and mean NEWS standardised per SD, age represented per year; p value for trend

In the whole cohort, high, but not low, delirium burden was associated with worse follow-up cognition in univariate and multivariate analyses: those experiencing 26 or more cumulative MDAS points had 0.6 SD deficit in follow-up cognitive scores (95%CI -0.94 to -0.26, $p<0.01$), compared with participants of similar baseline cognition, frailty and illness severity who did not experience any delirium (table 5.11, fig 5.5).

An interaction between baseline cognition and delirium burden was also demonstrated in the association with long term cognition. Those with the lowest baseline cognition had similar scores at follow-up regardless of delirium exposure (z-score -1.35, -1.43, -1.16 in none, low and high burden respectively). Individuals with high baseline cognition – those starting at +2.0 SD in z-score – had demonstrable decline even without delirium (z-score +1.17; absolute decline of 0.83). However, those experiencing high delirium burden had an even larger absolute decline (z-score -0.22; absolute decline of 2.22). Margins for each combination of delirium burden (none, low, high) with baseline cognition are shown in table 5.12.

A sensitivity analysis demonstrated no association between hospitalisation alone and cognitive decline (table 5.13).

Table 5.12: Margins between delirium burden in study and baseline cognition with follow-up cognition

Delirium burden	Baseline cognition	Margin	95% CI		P value
None	Low	-1.35	-1.50	-1.19	<0.01
None	Medium	-0.09	-0.14	-0.04	<0.01
None	High	1.17	1.04	1.31	<0.01
Low	Low	-1.43	-1.86	-1.01	<0.01
Low	Medium	-0.04	-0.26	0.18	0.72
Low	High	1.35	0.94	1.77	<0.01
High	Low	-1.16	-1.55	-0.77	<0.01
High	Medium	-0.69	-1.02	-0.36	<0.01
High	High	-0.22	-0.85	0.42	0.51

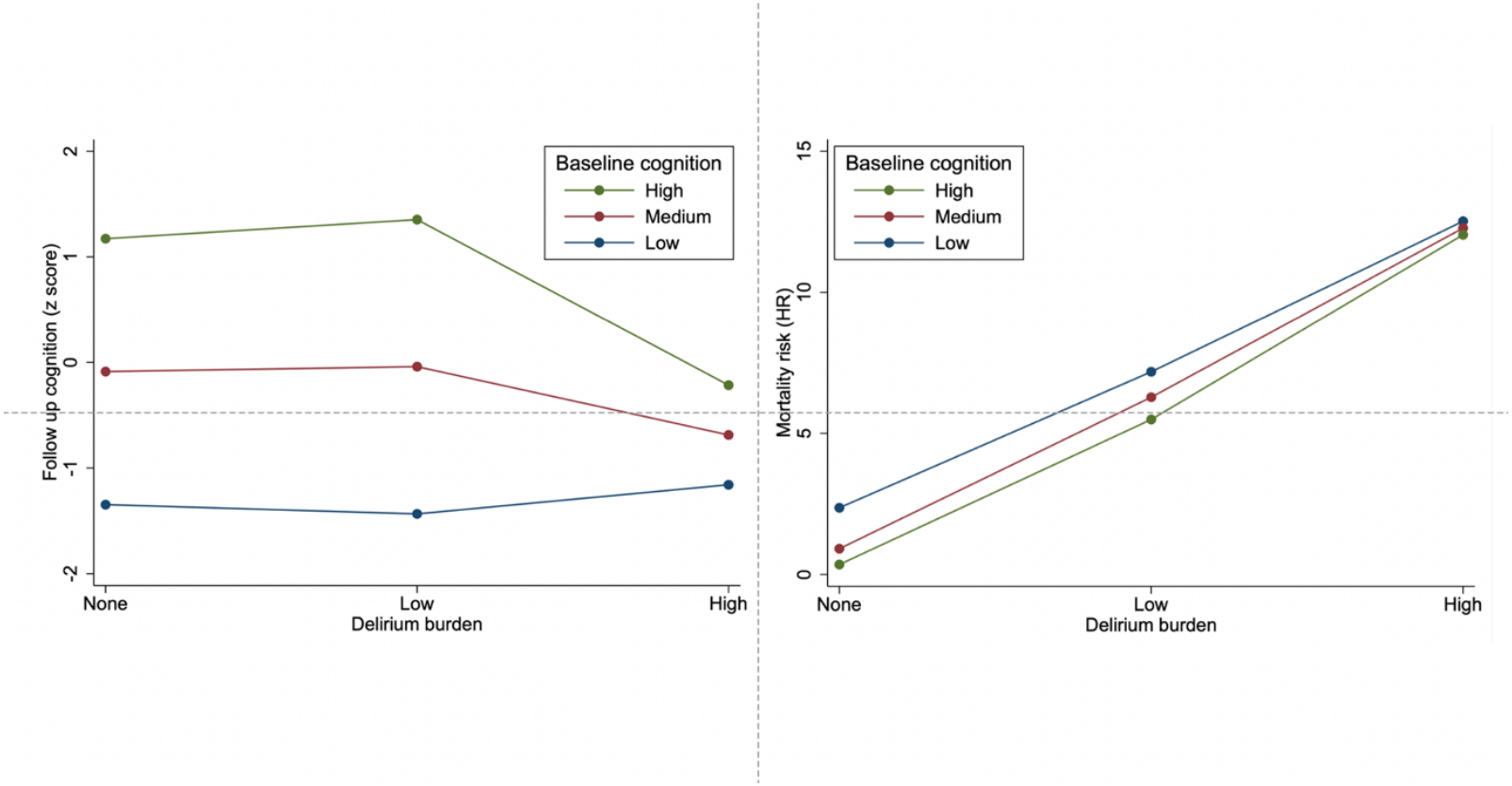
Marginal coefficients for effects of delirium burden and baseline cognition on follow up cognition, derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures. Baseline cognition divided into tertiles, delirium burden defined as none, low (under median cumulative MDAS burden <=26) and high (above median cumulative MDAS burden >26)

Table 5.13: Sensitivity analyses demonstrating no association between hospitalisation and follow up cognition

n = 1218	Multivariate Analyses			
	Beta	95%CI		P value
Hospitalisation	-0.02	-0.23	0.19	0.83
Baseline Cognition	0.63	0.56	0.69	<0.01
Hospitalisation Y/N - Baseline Cog interaction	-0.06	-0.20	0.07	0.36
Age	-0.02	-0.03	-0.01	<0.01
Female	0.05	-0.05	0.14	0.33
Frailty index	-0.08	-0.15	-0.02	0.02
NEWS	-0.02	-0.09	0.05	0.57

Linear regression analyses for sensitivity. Baseline and follow up cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; Hospitalisation defined as admission during study duration, FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score. Multivariable analyses show coefficients mutually adjusted for all other factors. Baseline cognition, FI and mean NEWS standardised per SD, age represented per year

Figure 5.5: Interactions in relationship between baseline cognition and delirium burden on long term cognition and mortality risk



Linear multivariable-adjusted estimates showing deleterious effects of high but not low delirium burden on long term cognition, with the biggest adverse impact on those with the best baseline cognition; Cox proportional hazards estimates, multivariable adjusted showing significant contribution of high delirium burden to mortality risk in medium and high baseline cognition groups but not low baseline cognition group

Left panel: z-score of follow up cognition (Data presented in Tables 8.12) *Right panel*: (Data presented in Tables 8.17). Baseline cognition defined by the modified Telephone Interview for Cognitive Status, augmented by two verbal fluency tasks. NEWS: National Early Warning Score. Baseline cognition, frailty index and NEWS standardised to show comparable effect sizes.

5.3.4.2 Mortality

In the 93 participants who died, these occurred within a median 444 days (IQR 282 to 747 days) of recruitment to the study. 63 of these deaths were in participants hospitalised at any stage during the study; 30 participants who died were not admitted during the study. Of the participants who died after experiencing hospitalisation, 18 had experienced low delirium burden and 45 high delirium burden.

Overall, delirium incidence was associated with increased mortality (HR 6.34, 95%CI 3.20 to 12.56, p<0.0001) (table 8.14) on full adjustment in multivariate analyses. Poorer baseline cognition, increasing age, male sex, increased frailty and increased illness severity were all also associated with increased mortality risk. While educational attainment was associated with increased mortality risk in univariate analysis, this association attenuated on full adjustment including baseline cognition. Delirium incidence during the study was associated with five times the increased mortality risk posed by each increasing year of age.

An interaction between baseline cognition and delirium incidence with mortality risk was demonstrated (HR 1.35 95%CI 1.00 to 1.82 p =0.05). Regardless of whether delirium was experienced or not during the study, mortality risk decreased with increasing baseline cognition. However, the hazard ratio for any patient who experienced delirium, regardless of baseline cognition, was greater than for those without delirium with the lowest baseline cognition (table 5.15)

Table 5.14: Associations between delirium incidence during study and mortality risk

	Univariate Analyses				Multivariate Analyses			
n = 1510	**HR**	**95% CI**		**P value**	**HR**	**95% CI**		**P value**
Delirium (Y/N)	15.14	10.06	22.78	<0.0001	6.34	3.20	12.56	<0.0001
Baseline Cognition*	0.58	0.51	0.64	<0.0001	0.71	0.56	0.91	0.01
Delirium (Y/N) - Baseline Cog interaction					1.33	0.98	1.80	0.07
Education**								
Degree level	Ref				Ref			
Up to secondary (12y schooling)	0.54	0.32	0.91		0.88	0.50	1.55	
Up to primary (6y schooling)	0.25	0.15	0.40	<0.0001	0.71	0.42	1.21	0.45
Age	1.11	1.08	1.14	<0.0001	1.05	1.01	1.08	0.01
Female	0.63	0.42	0.95	0.03	0.63	0.41	0.95	0.03
FI (minus cog)*	1.83	1.62	2.07	<0.0001	1.20	1.00	1.45	0.05
NEWS (mean)*	1.55	1.44	1.66	<0.0001	1.21	1.06	1.38	<0.0001

Linear regression analysis demonstrating association of delirium incidence during study with increased mortality risk. An interaction term is demonstrated between delirium incidence and baseline cognition.

Baseline and follow up cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score, calculated as mean over study admission days. Multivariable analyses show coefficients mutually adjusted for all other factors.

*Baseline cognition, FI and mean NEWS standardised per SD, age represented per year **p value for trend

Table 5.15: Margins for delirium prevalence, baseline cognition and mortality risk

Delirium incidence	Baseline cognition	Margin	95% CI		P value
No	Low	54.38	-98.79	207.55	0.49
No	Medium	27.68	-50.50	105.86	0.49
No	High	14.09	-26.97	55.15	0.50
Yes	Low	195.66	-345.76	737.08	0.48
Yes	Medium	175.42	-323.26	674.10	0.49
Yes	High	157.27	-313.14	627.69	0.51

Marginal coefficients for effects of delirium prevalence and baseline cognition on mortality risk, demonstrating the greatest mortality risk occurs in those who experienced delirium with poorest baseline cognition. Baseline cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures. Baseline cognition divided into tertiles.

Increasing delirium burden was associated with increased mortality risk on full adjustment in multivariate analyses (HR 7.14 95% CI 3.50 to 14.58 for low delirium burden; HR 13.74 95% CI 6.75 to 27.99 for high delirium burden respectively, p <0.0001 for trend) (table 8.16).

An interaction between delirium burden and baseline cognition with mortality risk was also demonstrated in multivariate analyses. Poorer baseline cognition was associated with increased risk of death for participants who experience no or low burden of delirium. Participants who experienced low delirium burden, regardless of baseline cognition, had a greater mortality risk than those without delirium but also the poorest baseline cognition. However, for participants with the greatest delirium burden, mortality risk was similar regardless of baseline cognition, with hazard ratios of 9.75, 10.02 and 10.29 for low, medium and high baseline cognition respectively. Margins for each combination of delirium burden (none, low, high) with baseline cognition towards mortality risk are shown in table 5.17. The differential mortality risks stratified by baseline cognitive status in participants with no or low delirium burden, but not demonstrated in those with high delirium burden, is shown in figure 5.6.

Table 5.16: Associations between delirium burden and mortality risk

		Univariate Analyses				Multivariate Analyses (Groups as continuous)			
n = 1510		HR	95% CI		P value	HR	95% CI		P value
Delirium Burden	None	Ref				Ref			
	Low	8.63	4.81	15.49		7.14	3.50	14.58	
	High	24.13	15.18	38.37	<0.0001	13.74	6.75	27.99	<0.0001
Baseline Cognition*		0.58	0.51	0.64	<0.0001	0.64	0.49	0.83	<0.0001
Delirium Burden - Baseline Cog interaction	None					Ref			
	Low					1.53	0.93	2.51	
	High					1.59	1.14	2.20	0.02
Education**									
Degree level		Ref				Ref			
Up to secondary (12y schooling)		0.54	0.32	0.91		1.00	0.57	1.74	
Up to primary (6y schooling)		0.25	0.15	0.40	<0.0001	0.71	0.42	1.20	0.35
Age		1.11	1.08	1.14	<0.0001	1.27	1.02	1.57	0.03
Female		0.63	0.42	0.95	0.03	0.66	0.44	1.01	0.06
FI (minus cog)*		1.83	1.62	2.07	<0.0001	1.18	0.98	1.42	0.09
NEWS (mean)*		1.55	1.44	1.66	<0.0001	1.13	0.99	1.29	0.06

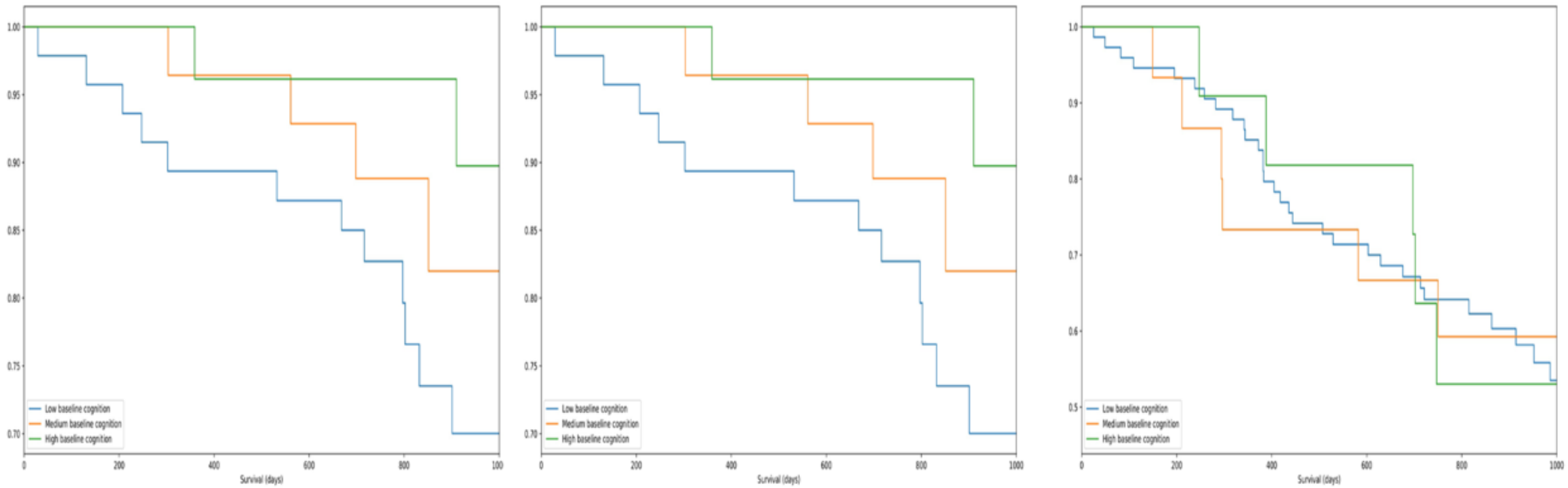
Linear regression analyses. Baseline and follow up cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures; FI frailty index, minus cognitive items to avoid collinearity; NEWS national early warning score, calculated as mean over study admission days. Delirium burden derived by thresholding cumulative MDAS scores of participants who experienced delirium at median. Multivariable analyses show coefficients mutually adjusted for all other factors. Baseline cognition, FI and mean NEWS standardised per SD, age represented per year; * Variable centred and standardised ** p value for trend

Table 5.17: Margins for delirium burden, baseline cognition and mortality risk

Delirium burden	Baseline cognition	Margin	95% CI		P value
No	Low	1.78	0.52	3.05	0.01
No	Medium	0.73	0.40	1.06	<0.0001
No	High	0.30	0.10	0.50	<0.0001
Low	Low	5.45	-0.65	11.56	0.08
Low	Medium	5.21	0.95	9.46	0.02
Low	High	4.97	-1.57	11.50	0.14
High	Low	9.75	2.03	17.47	0.01
High	Medium	10.02	1.54	18.49	0.02
High	High	10.29	-1.26	21.84	0.08

Marginal coefficients for effects of delirium burden and baseline cognition on mortality risk, with the greatest cognitive decline demonstrated by those with the highest deliirum burden and baseline cognitive performance. Baseline cognition derived from modified Telephone Interview of Cognitive Status plus two verbal fluency measures. Delirium burden, defined as no, low (under median cumulative MDAS burden <=26) and high (above median cumulative MDAS burden >26). Baseline cognition divided into tertiles.

Fig 5.6: Kaplan Meier plots by delirium burden and baseline cognition



Protective effects of high and medium baseline cognition on mortality risk if exposed to low delirium burden or no delirium. Participants with exposure to high delirium burden experienced similar survival trajectories, regardless of baseline cognition as evidenced by overlapping survival lines.

5.4 Discussion

5.4.1 Summary of findings

These findings demonstrate that across this population sample, higher baseline cognition reduces the risk of incident delirium. Over the whole population, in those with higher baseline cognition, any experienced delirium is of shorter duration, even after accounting for age, illness severity, educational attainment and frailty. However, the relationship between baseline cognition and delirium severity is non-linear: if delirium presents during acute illness, participants with the lowest and highest cognitive baselines experience the most severe delirium and arousal abnormalities. Delirium severity and abnormal arousal were closely related at all levels of cognition.

Overall, delirium incidence increased cognitive decline and risk of death in a dose dependent manner at long-term follow up, as demonstrated by the association of cognitive decline with increasing delirium burden. However, the protective effects of higher baseline cognition on cognitive decline and future mortality risk only seemed apparent to those who experienced no delirium or low delirium burden. For participants who experienced high delirium burden, rates of cognitive decline and mortality risk were similar, suggesting an attenuation of the protective effects of baseline cognition in patients experiencing high delirium burden.

Taken together, when acute illness is sufficient to lead to delirium, different factors may be at play across the range of baseline cognitive function. In the context of normal baseline cognition, recognising delirium could be the strongest indication of acute illness in older people, over and above physiological indices such as NEWS. These findings highlight how maximising baseline cognition can minimise later life cognitive decline and mortality risk. Moreover, the toxic effects of high delirium burden appear to overcome these protective effects of baseline cognition, emphasising the importance of preventing or reducing delirium severity during acute illness in older people.

5.4.2 Significance and reasons for findings

The effect size of delirium on cognitive decline is substantial, comparable to that observed in hereditary forms of Alzheimer's disease. DELPHIC participants with the highest baseline cognition declined by 0.83 and 2.22 SDs with 'no' and 'high' delirium burden exposure respectively after two years on full adjustment. This is worse than that observed in the highest risk groups in the Alzheimer's Dementia Neuroimaging Initiative (ADNI), who are positive for amyloid in PET imaging or in CSF samples and demonstrated declines of 1 SD over approximately three years (Mormino,

Papp et al. 2017), as well as to amyloid related decline in the normal population, estimated to be 0.15 to 0.35 SD per year (van der Kall, Truong et al. 2021). My analysis suggest delirium *per se* is neurotoxic resulting in increased cognitive decline and increased mortality, likely via systemic dysfunction. Although the disorientating and distressing effects of inpatient hospitalisation are well recognised, the effects of delirium on cognitive decline are independent of admission to hospital.

Limitations in illness severity ascertainment could confound the apparent finding that delirium is more adverse in those with high baseline cognition. Is it possible that those with higher baseline cognition were also less frail to begin with, hence the threshold for hospitalisation was higher for this subgroup and is a marker of higher illness severity not captured by NEWS alone? Conversely, was there a floor effect in detecting further cognitive decline on follow up in those with the poorest cognitive performance at the start of the study? This possibility is less likely because the baseline cognitive scores are normally distributed, with no apparent floor effects.

Why might individuals with better baseline cognition be more vulnerable to cognitive decline associated with delirium? It is possible that the physiological precipitants of delirium in these individuals somehow had a more direct neurological impact than in those with worse baseline cognition. This finding echoes a previous result of a disproportionate impact of delirium on fitter individuals (Dani, Owen et al. 2018) or those without prior dementia (Pitkala, Laurila et al. 2005). In this context, delirium could be the best marker of acute illness severity in those with highest baseline cognition. Pragmatically, delirium should raise concern for subsequent cognitive decline in those with the “most to lose”. Cumulative MDAS scores are associated with longer length of stay, which may be a proxy for either a more severe illness severity or poorer functional baseline, even beyond what can be accounted for by NEWS and frailty index, respectively. As a result, compared with NEWS, cumulative MDAS burden across an admission may be a more informative metric.

5.4.3 Alignment with current literature

5.4.3.1 Poor baseline cognition with increased risk of delirium

These broader findings align with previous experimental data from animal studies showing higher grades of prior neurodegeneration result in more severe and more prolonged delirium symptoms when challenged with a standardised inflammatory stimulus. In APP/PS1 mice, microglia were found to produce an exaggerated interleukin 1 beta (IL-1 beta) response and subsequent exaggerated astrocyte chemokine response to lipopolysaccharide-induced IL1-beta. This resulted in greater amyloid deposition and neuroinflammation. Macroscopically, hippocampal networks in

APP/PS1 mice were primed and hypersensitive to IL-1 beta challenge, resulting in an acute cognitively dysfunctional state analogous to delirium (Lopez-Rodriguez, Hennessy et al. 2021).

Similar associations between premorbid cognitive impairment with delirium risk are evident in epidemiological studies. Any baseline cognitive impairment (defined as pre-diagnosed dementia, MMSE<23 or 24, IQCODE positive or TICS<30) was associated with an odds ratio for delirium between 1.3 and 5.5 (Fong, Davis et al. 2015). In a sample of the oldest-old, the odds ratio for incident dementia with delirium was 8.7 (Davis, Muniz Terrera et al. 2012). The findings that there is a non-linear relationship between baseline cognition and delirium severity, with those at lowest and highest cognitive baselines being at risk of the most severe delirium, is novel and extends current knowledge.

In addition to increasing age, as anticipated, being related to increasing cognitive decline, the multivariate models also highlight frailty as an independent risk factor for worse cognitive impairment at follow up. This is consistent with previous studies that demonstrated frailty as a modifier for the clinical expression of dementia neuropathology, with frailer participants more likely to demonstrate an association between burden of Alzheimer's neuropathological burden and expression of Alzheimer's dementia symptoms (Wallace, Theou et al. 2019).

5.4.3.2 Delirium exposure produces differential cognitive and mortality risks depending on baseline cognition

The results are consistent with previous findings that prior delirium exposure is a risk factor for cognitive decline, independent of classical neuropathologies associated with dementia (Davis, Muniz Terrera et al. 2012). These findings support delirium to be directly toxic towards cognitive decline, instead of being an epiphenomenon of dementia during acute illness.

The finding that delirium is an independent risk factor for mortality is also consistent with an earlier meta-analysis (Witlox, Eurelings et al. 2010). However, I note that the hazard ratios demonstrated in this cohort was over three times higher than previously reported in meta-analyses (6.5 vs 1.95). It is likely that more robust adjustment within our population cohort, for confounders not necessarily accurately possible in a meta-analysis such as baseline cognition, illness severity and baseline frailty, produced a more accurate hazard estimate. The finding that the adverse effects of increased delirium exposure is stratified by baseline exposure, in particular those with the best cognitive baseline have the most to lose after delirium, is a novel extension to current knowledge.

5.4.4 Strengths and limitations

5.4.4.1 Limitations

My findings should be interpreted in the context of several limitations. While recruited participants closely matched the sampling frame by age and socioeconomic position, the relative overall response rate to initial invitation was low, and specific reasons for non-participation remain unknown. Segments of the Camden population, for whom English is not their first language, are ethnically non-White Caucasian, are communicatively non-verbal or are unable to consent independently but lack a nominated next of kin or formal advocate, were under-represented.

Second, although I had comprehensive methods to identify hospitalised participants, there is inevitably a degree of selection bias towards inpatients that would have missed cases who developed delirium but remained in the community. Furthermore, many patients were discharged with a confusional state that would still meet DSM-IV criteria for delirium on their final day of their inpatient stay. As a result, delirium duration could not be accurately defined for these patients while their true delirium burden would be under-estimated as their clinical course continued into the community.

Despite the advantage of frequent clinical assessments, I made assumptions about missing data on delirium status over weekends and public holidays. I pragmaticaly decided that the nature of these missing data was not sufficient completely at random and that the employed approach of alternate forward and backfilling from the last ascertained date would be a reasonable compromise.

My analyses did not account for any medication-related effects, nor did I explore possible differences attributable to underlying aetiology; these are areas of ongoing analyses. Examining delirium in further granularity would be of particular interest in potentially stratifying patients to differing management approaches with prognostication implications. Although I used validated inpatient scales for physiological derangements, these may not adequately quantify illness severity that particularly in oldest-old patients. There was appreciable loss to follow-up, though because this was more likely in those with poorer baseline cognition, my models may have under-estimated the impact of delirium on cognitive outcomes. In common with other observational studies, model estimates are subject to residual confounding.

5.4.4.2 Strengths

The novel design of the DELPHIC study and its unique position to answer questions regarding the incidence and adverse deleterious consequences of delirium have been well described. Specifically, prospective capture of brain symptoms before and during acute illness allows for the most rigorous mapping of baseline cognition, hospitalisation and delirium in a community sample to date. DELPHIC made robust ascertainments of baseline and follow up cognition, function and co-morbidities. Our clinical features have been designed to allow direct comparisons to current DSM-IV delirium criteria, while the use of validated scales such as MDAS in full or in part allow sensitivity analyses of cortical and subcortical elements, hence examining delirium specific effects compared to those possibly caused by premorbid cognitive impairment.

Although single studies are infrequently sufficiently powered for analyses with multiple exposures and multiple outcomes, all minimum sample sizes were calculated and published as a protocol prior to the start of DELPHIC. While delirium was specified as the primary outcome in the analyses of association between baseline cognition and delirium phenomenology, prior to DELPHIC, there was no consensus within literature on how best to operationalise "delirium", whether as incidence, duration or a measure of cumulative exposure burden. However, the richness and size of the DELPHIC dataset, a main strength of this thesis, allowed multiple delirium definitions to be constructed. Regardless of how delirium was operationalised, the findings were consistent, further demonstrating the robustness of this thesis' conclusions. The current terms to operationalise delirium within DELPHIC were arrived at via iterative discussions among contributors: it is envisaged these terms may further evolve, for even the same dataset, as more advanced methods are applied in future analyses. Similar progress can be anticipated within the delirium field, leading to greater consensus of what would constitute standardised delirium measures in future research.

5.4.5 Summary and future directions

Overall, these findings translate to significant clinical implications. It is vitally important to ascertain a patient's baseline cognition on first presentation to stratify the subsequent increased risk of significant cognitive decline and death. Baseline cognition also helps identify patients most likely to experience the most severe delirium, as well as being those with "the most to lose" cognitively if faced with prolonged delirium exposure, emphasising the particular efforts needed to manage underlying precipitating and perpetuating factors in these specifically high risk cohorts.

My findings open the possibility for offering more accurate prognostication of both delirium risk during future acute insults, informing patient expectations when making decisions on elective

surgeries and unexpected illnesses. For example, patients with baseline cognitive function in highest are likely to experience severe delirium, should this occur, which may be informative when consenting for general anaesthesia during an elective operation. We can also now prognosticate the likelihood of adverse outcomes depending on delirium duration: in an analogous example, the same patients with higher baseline cognition should also be informed that should they experience high delirium burden, despite being at lower risk of exposure, they will disproportionately vulnerable to subsequent cognitive decline, experiencing more than two standard deviations of decline in general cognition over two years.

Lastly, these findings highlight the complex interactions between baseline cognition and acute illness. It is important to understand in even greater granularity if further nuanced associations exist when other modalities are taken into account, for example after inclusion of clinical phenotype, biochemistry, neuroimaging. Only then can personalised pathways of management and recovery measures be best targeted to the most appropriate patients.

6 Individualised patient prediction of mortality at 600 days post-discharge

Chapter Outline

- Introduction
- Methods
  - Kaplan Meier and Cox proportional hazards
  - XG boost
  - Anatomical inference
- Results
  - Multimodal prediction of mortality
  - Anatomical correlates of mortality and predictive performance
- Discussion
  - Predictability of mortality from routine data
  - Mechanisms
  - A multimodal index of frailty

Chapter 5 demonstrated the complex non-linear relationships between multiple contributing factors that lead to adverse outcomes across a population. However, our aspiration of personalised medicine requires translation of population associations into individualised prediction models.

The second project of this thesis aims to offer a proof of concept for deploying multimodal high dimensional machine learning prediction algorithms for predicting outcomes in older patients with multimorbidities. I will use mortality as the predicted outcome for several reasons: first, it is the most accurately ascertained outcome in older people. Second, prognosis is one of the most common questions asked during acute illness. In addition, clinically, robust mortality prediction is critical for healthcare resource prioritisation and appropriate alignment of clinical interventions to patient prognoses (Powers, Chaguturu et al. 2015). In the context of this thesis, mortality risk provides an optimal foundation to build a pipeline of future prediction models for older people with multiple complex, interacting morbidities. The findings from this chapter were pubublished in a manuscript entitled “Predicting mortality in acutely hospitalised older patients” (Tsui, Tudosiu et al. 2023).

6.1 Chapter Introduction

As a clinical outcome of arguably the greatest concern, mortality prediction instruments should have the highest fidelity: to guide expectations, target interventions, and identify modifiable mechanisms of disease (Powers, Chaguturu et al. 2015). Though death is narrowed to specific causes, it is also a general, constitutional risk in older people, distributed across a wide field of biological and pathological factors, both incident and enduring. The determinants of such vulnerability may be less in any specific condition than in the complex interaction of multiple accruing co-morbidities and age-related physiological changes that single disease-centred models cannot satisfactorily capture. Predicting mortality at older ages should require a patient-centred, fully-inclusive, population-focused approach, capable of absorbing the wide heterogeneity of factors plausibly determining individual risk: older people are the largest contingent of healthcare users, with the most variable intragroup functional and cognitive performance ranges (Lowsky, Olshansky et al. 2014). Though making up only 18% of the UK population, patients aged over 65 account for 42% of acute hospital admissions ((ONS) 2018), a gap projected to widen as those aged over 60 double in number by 2050 worldwide .

Causes of mortality after an acute hospital admission are heterogeneous and inherently challenging to predict. Chronological age alone has long been recognised to be poorly predictive of survival, particularly in older people (Knaus, Wagner et al. 1991). Several short and medium-term mortality prediction instruments exist, such as APACHE-III (Knaus, Wagner et al. 1991), HELP (Teno, Harrell et al. 2000), BISEP (Inouye, Bogardus et al. 2003), SAFES (Drame, Jovenin et al. 2008) and HOMR (van Walraven 2014). However, their applicability to acutely admitted unselected older patients is limited by lack of validation beyond specific clinical settings such as critical care (Knaus, Wagner et al. 1991), poor calibration with age (Fischer, Gozansky et al. 2006), variation in performance across population groups (Curtin, Dahly et al. 2019), or dependence on background information not readily available in the acute setting (van Walraven 2014). All are linear constructions from relatively limited input variables, unlikely sufficient for accurately modelling the complex processes determining mortality risk after hospital admission: they do not take into account cognitive status or neuroimaging, despite established associations between dementia, delirium and later-life mortality in population studies (Tsui, Searle et al. 2022). They do not utilise direct or surrogate measures of musculoskeletal veracity such as soft tissue or bone density, despite associations between sarcopenia and osteoporosis with increased mortality risk via low-energy traumatic fractures.

Current instruments fail to address two cardinal potential characteristics of the problem: the distribution of factors relevant to mortality across *multiple* clinical domains and investigational modalities, and the likely presence of heterogeneous causal interactions accessible only to *complex*, high-dimensional statistical models. No instrument attempts to draw power from the synthesis of multiple, individually weakly predictive features—clinical or investigational—that may in aggregate be both highly predictive and robust to the distributional heterogeneities commonly observed in real-world healthcare data. Identifying mortality risk distributed across multiple investigational modalities, driven by non-linear interactions between remote variables, remains unexploited.

Recent advances in complex modelling now allow a radically different approach. We can move beyond simple, unimodal, low-dimensional models to complex, multimodal, high-dimensional models that integrate rich information acquired during routine care. Crucially, we can *quantify* the comparative benefit—evaluated on out-of-sample data—of increasing model complexity to show the optimal prediction strategy. If higher model complexity is beneficial, an efficient strategy would be to vastly increase the number of model parameters using existing, routinely collected clinical data. When mortality prediction becomes saturated, further improvement would likely require the introduction of biologically novel measures not routinely captured. The distinction matters because the former can be rapidly introduced with algorithmic innovation only, without disturbing care pathways, where the latter is constrained by the long timelines and need to validate novel clinical investigations.

I used a large, consecutive, unselected, fully-inclusive cohort of older patients acutely admitted to hospital to a) quantify the predictability of 2-year mortality from routinely acquired multimodal clinical data, b) compare the performance of predictive models varying in input modality, dimensionality and architectural flexibility, c) identify candidate causal mechanisms of increased mortality, and d) establish the foundations of a readily deployable clinical tool for predicting all-cause mortality in unselected older patients admitted to acute hospitals, to be fully developed in future large-scale, multi-centre studies.

6.2 Methods

The dataset used for chapter 6 was described in chapter 4.4.2. The methods for constructing survival analyses, machine learning prediction models and neuroanatomical inferences were described in chapter 4. In summary, for the prediction models utilized in this chapter, a XGboost classifier was trained using ten-fold cross validation and hyperparameters optimised using grid searches. Performance was defined as area under the curve receiver operator characteristic

(AUCROC). Segmented brain sequences were linearly compared between ground truth and predicted alive-dead contrasts using SPM.

6.2.1 Kaplan Meier and Cox Proportional Hazards

I calculated length of time elapsed between each patient episode until the end of study data collection, categorising survival status at 600 days after initial hospital admission. This threshold was chosen based on approximately half of patients were alive at this timepoint. If a patient remained alive at each survival point, or if time elapsed was fewer than 1200 days from hospital admission, their data was censored. I plotted a Kaplan-Meier curve to illustrate survival up to 1200 days after initial hospital admission, separated by each acute cognitive status class. Next, a Cox proportional hazards model was estimated using acute cognitive status, age, sex. For these initial models, I also adjusted for mean red cell distribution width, mean CRP, and mean baseline creatinine, because these are good clinically accepted markers for acute illness and biological indicators of baseline morbidity.

6.2.2 XG boost classifier model

I constructed a predictive model for 600 days post-discharge mortality with the gradient-boosting machines-based algorithm XGBoost (Chen T 2016). The choice of algorithm was motivated by the combination of robustness, flexibility, data efficiency, and optimisability given the scale of available data. To quantify the value of increased dimensionality, I estimated an array of models incrementally increasing in number and range of input variables : 1) age and sex (2 variables); 2) primary diagnosis, age and sex (17 variables); 3) cognitive status, age and sex (4 variables); 4) primary diagnosis, cognitive status, age and sex (19 variables); 5) bloods, primary diagnosis, cognitive status, age and sex (91 variables); 6) CT intracranial, primary diagnosis cognitive status, age and sex (5367 variables); 7) CT extracranial, primary diagnosis, cognitive status, age and sex (12989 variables); 8) CT whole brain, age and sex (18399 variables); 9) CT whole brain, bloods, primary diagnosis, cognitive status, age and sex (18494 variables); 10) CT whole brain, primary diagnosis, cognitive status, age and sex (18422 variables). The target outcome for all models was survival at 600 days from admission.

The data were randomly split into training (70%) and testing (30%) partitions, stratified by 600-day mortality outcome to avoid uninformative training subsets without patients who had died. Where multiple CT images were obtained in the same admission episode, the first image was always used. The test partition contained unique patients only, which ensures avoidance of performance inflation. XGB models were trained and optimised using ten-fold cross-validation from the training

partition with 600-day mortality and the AUCROC as the evaluation metric. A manually targeted grid search followed a random initial parameter grid search to optimise model hyperparameters (number of estimators, maximum depth, minimum child weight, learning rate, gamma, subsample, column sample by tree. The best performing fold hyperparameters as defined by maximal AUCROC were used to quantify performance on held-out test data, evaluated through ten-fold cross-validation of the test set.

6.2.3 Anatomical inference

To understand the anatomical patterns driving the imaging contribution to model fidelity, I sought to identify linear and non-linear voxel-wise associations with the target outcome. First, to quantify the sensitivity of the imaging data to known anatomical variations with dementia status, I performed standard voxel-wise mass-univariate volumetric brain morphometry with grey and white matter tissue compartments as the input using CTseg function on SPM. Next, I used the same approach to identify linear anatomical associations with mortality separately within grey matter, white matter, soft tissue and bone tissue compartments, adjusted for age, sex, delirium status, dementia status, total brain volume and degree of atrophy. Third, to highlight potentially non-linear associations captured by the XGB model, its feature importances, indexed by ranked Gini impurity, were projected back into MNI space for anatomical visualisation.

6.3 Results

6.3.1 Study summary

2951 admissions episodes with cognitive status and primary diagnoses were recorded from 1855 unique patients during the study period. Following grouping of admissions within 28 days as a single episode, 1975 admission episodes, from 1601 unique patients, were defined with linkage to at least one complete set of blood tests within the first 48 hours of admission. Last, in the whole cohort, 804 admission episodes could be linked to a CT head 28 days before or after the day of admission. The mean age of the complete cohort was 84.5 years (table 6.1). Patient cognitive status (within an admission episode) were as follows: 44% were cognitively intact, 16% had a diagnosis of dementia alone, 21.9% delirium alone and 18% delirium superimposed on dementia; 36.6% of admission episodes resulted in death 600 days after their day of admission (figure 6.1). The proportion of missing bloods data is described in table 6.2.

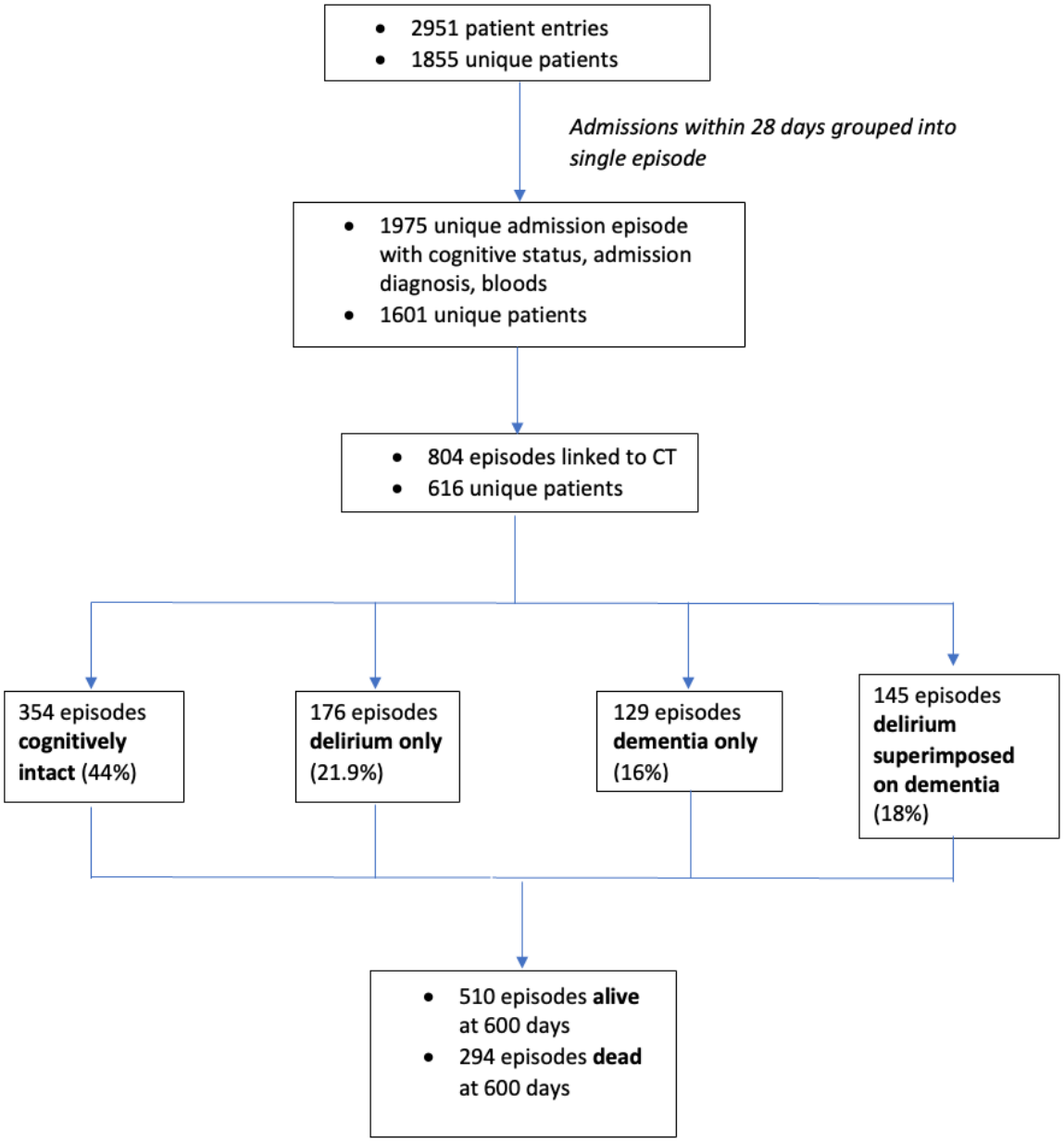


Fig 6.1 – Flow chart of patient inclusion and complete cohort compilation

Table 6.1 – Descriptive summary of complete cohort

	Alive at 600 days (n = 510)	Dead at 600 days (n = 294)	p value*
Age (mean, SD)	83.7 (7.0)	85.9 (6.9)	<0.01
Women (%)	27.60%	16.80%	<0.01
Cognitive status			
Cognitively intact (%)	46.70%	39.77%	
Delirium only (%)	20.35%	24.86%	
Dementia only (%)	17.03%	14.36%	
Delirium superimposed on dementia (%)	15.93%	22.10%	0.02**
Total brain volume (mls)	917 (89)	908 (98)	0.2
Atrophy (%)	81.4 (2.1)	81.1 (2.0)	0.35

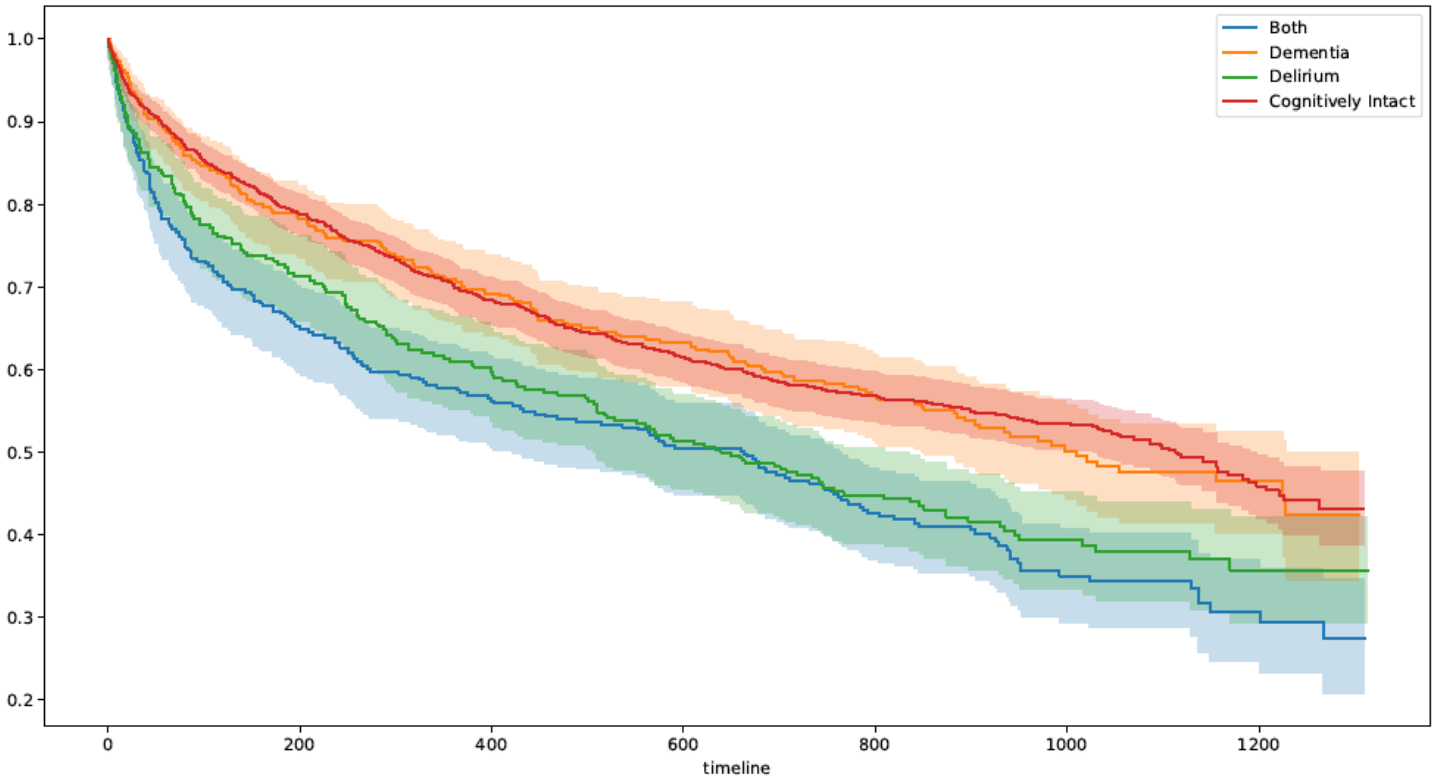
Mean values for continuous variables (standard deviation in brackets). *Tests for significance: chi squared for categorical variables, 2 sample t-test for continuous variables. **p value for trend, assuming hierarchical increase in cognitive decline.

Table 6.2 – Percentage of blood data missing from original dataset

Blood	Percentage Missing
Corrected Calcium	25.22%
Prothrombin Time	14.73%
Alanine Aminotransferase	9.42%
Bilirubin	8.30%
Alkaline Phosphatase	7.95%
Albumin	6.68%
Potassium	4.15%
C-Reactive Protein	1.27%
Red Cell Distribution Width	0.66%
Platelet	0.51%
Creatinine	0.51%
Haemoglobin	0.46%
Monocyte	0.46%
Mean Corpuscular Volume	0.46%
Basophil	0.46%
Eosinophil	0.46%
Red Cell Count	0.46%
Lymphocyte	0.46%
Haematocrit	0.46%
White Cell Count	0.46%
Neutrophil	0.46%
Urea	0.35%

At 600 days after the admission date, 51% of patients in the complete cohort were alive. Crude survival of individual patients, stratified by cognitive status, demonstrated increased mortality risk in patients with delirium and delirium superimposed on dementia (figure 6.2).

Fig 6.2: Unadjusted Kaplan Meier plot of survival as stratified by cognitive status during acute illness



Multivariate Cox proportional hazards demonstrated increased mortality risk with increasing age, higher mean CRP and red cell distribution width (table 6.3). While cognitive status alone was not significant in univariate analysis, a trend for increased mortality with worsening cognitive status was significant on full adjustment with covariates including degree of brain atrophy.

Table 6.3: Univariate and multivariate Cox proportional hazards model

	Univariate				Multivariate			
	HR	**95% CI**		**p value**	**HR**	**95% CI**		**p value**
Age	1.04	1.03	1.06	<0.01	1.04	1.03	1.06	<0.01
Sex (Male)	1.34	1.11	1.62	<0.01	1.22	0.99	1.50	0.07
Cognitive status*								
Cognitive intact	Ref				Ref			
Delirium	1.15	0.90			1.02	0.79	1.32	
Delirium and Dementia	1.11	0.85			1.25	0.94	1.65	
Dementia	1.34	1.04	0.16	0.16	1.45	1.12	1.89	0.02
CRP (mean)**	1.36	1.25	1.48	<0.01	1.27	1.14	1.40	<0.01
RDW mean**	0.74	0.67	0.81	<0.01	0.81	0.73	0.90	<0.01
Creatinine (mean)**	0.78	0.71	0.86	<0.01	0.90	0.81	1.01	0.06
WCC (mean)**	1.19	1.07	1.31	<0.01	1.04	0.93	1.16	0.49
Atrophy**	0.96	0.87	1.05	0.39	1.07	0.96	1.18	0.21

Model full adjusted; HR hazard ratio

*p value for trend

**Standardised per standard deviation

6.3.2 Mortality prediction from multimodal data

Baseline XGB predictive models of age and sex achieved only a modest AUCROC of 0.672 (SD=0.0781) on out-of-sample test data (figure 6.3). There was no significant benefit from the addition of cognitive status (AUROC=0.677, SD=0.0654, p=0.631,) or primary diagnosis (AUCROC=0.673, p=0.216, SD=0.0643) alone. However, addition of both cognitive status and admission diagnosis (AUROC=0.698, SD=0.0649, p<0.001) and bloods tests to the full clinical model yielded a modest but statistically significant improvement over the baseline (AUCROC=0.692, p<0.001, 0.0479) that was not explicable by the intention to investigate alone.

A substantial jump in predictive performance was observed with the addition of imaging data to the full clinical model. The combination of intracranial imaging with demographic and clinical features yielded an AUROC of 0.82 (SD=0.0621), significantly higher than both baseline (p<0.001) and demographic and clinical data alone (p<0.001). The use of extracranial features produced marginally better performance (AUROC=0.848, p=0.0395, SD=0.0271) than intracranial features in otherwise identically specified models. Intracranial and extracranial compartments taken together with demographics advanced performance further (AUC = 0.854, p<0.001, SD=0.0484). The highest-performing model included the widest selection of inputs—demographics, clinical features, blood tests, intracranial and extracranial CT, exhibited an AUROC of 0.874 (SD = 0.455) (fig 6.4). Though the mean value was higher, this was not significantly different from the preceding model. Cross-validated training and test values are presented in table 6.5. The prediction model was re-applied onto the original population to demonstrate superior performance achieved by using bloods and neuroimaging compared to using acute cognitive status alone, illustrated in a Kaplan Meier plot showing predictive status alongside acute cognitive status (fig 6.5).

In optimising the XGboost model, grid searches of hyperparameters, using AUCROC as the optimising metric, was performed using pre-selected tuning parameters within the ranges in table 6.4a. The most optimal final hyperparameters listed in table 6.4b.

Table 6.4a: Hyperparameter 10 fold cross validation grid search range; **6.4b**: XGboost final hyperparameters

a) Hyperparameter grid search range

Hyperparameter	Grid search range
Early stopping rounds	200
n_estimators	Free
Maximum depth	[3, 5, 7, 9]
Minimum child weight	[1, 3, 5, 7, 9]
Gamma	0 to 1 (in 0.1 intervals)
Subsample	0.6 to 1 (in 0.1 intervals)
Column sample by tree	0.4 to 1 (in 0.1 intervals)

b) XGboost final hyperparameters

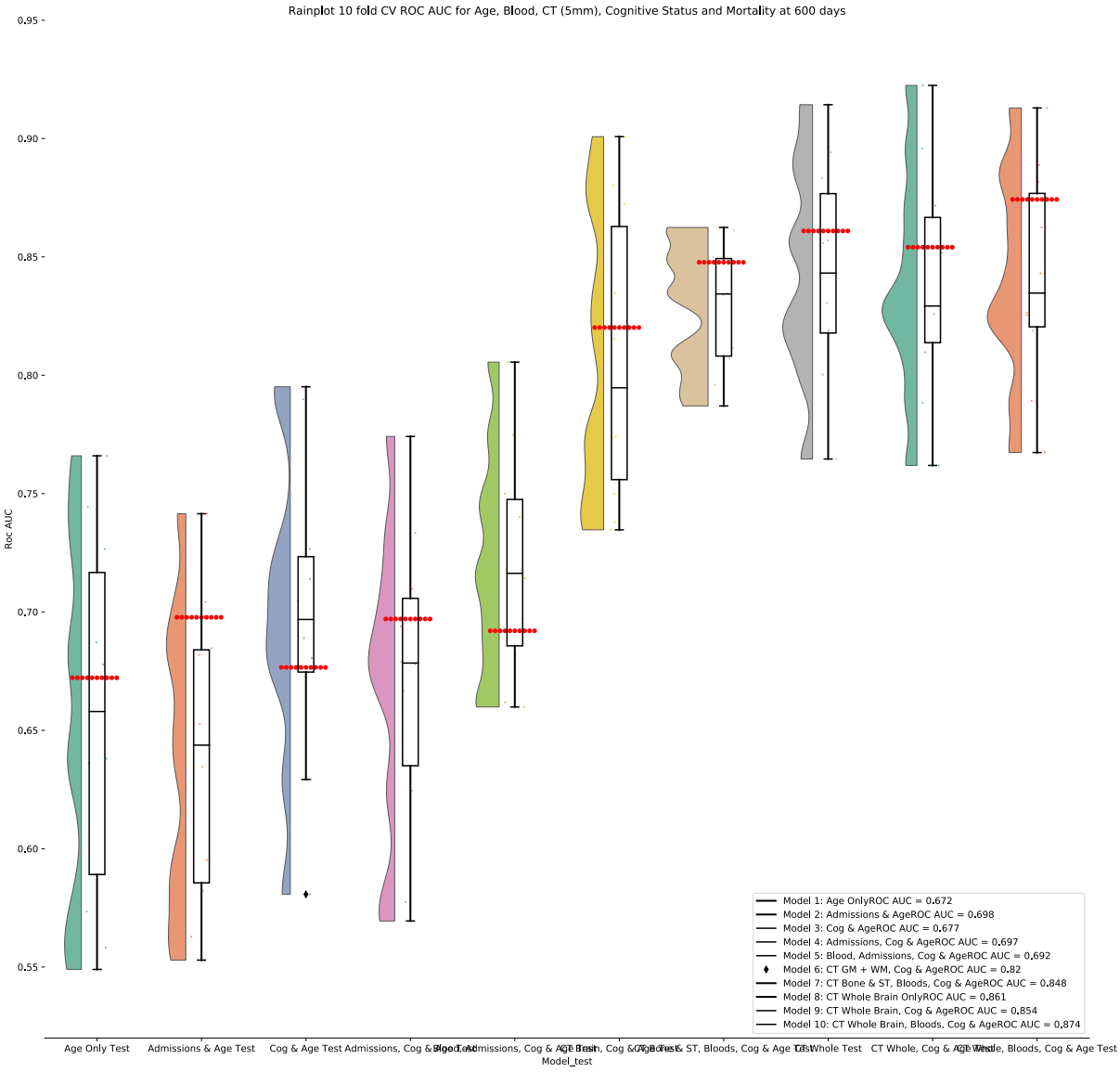
Hyperparameter	Value
Learning rate	0.1
Minimum child weight	1
N estimators	196
Maximum depth	7
Gamma	0
Subsample	0.9
Colsample by tree	0.6
Scale position weight	1
Random state	42

Table 6.5: 10 fold cross validation by hierarchical increasing number of modalities as predictive features

CV fold	Age & sex		Admissions, age & sex		Cognitive status, age & sex		Cognitive status, admissions, age & sex		Blood, cognitive status, admissions, age & sex	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test
1	0.573413		0.55291		0.580688		0.577381		0.714286	
2	0.687169		0.681878		0.680556		0.67791		0.681878	
3	0.558201		0.562831		0.672619		0.569444		0.69709	
4	0.677778		0.652778		0.688889		0.666667		0.661806	
5	0.765972	0.672191687	0.684722	0.697741326	0.795139	0.676571625	0.733333	0.697054277	0.75	0.69203023
6	0.744444		0.704167		0.713889		0.709722		0.805556	
7	0.636054		0.634694		0.789796		0.693878		0.740136	
8	0.638095		0.595238		0.704762		0.678912		0.718367	
9	0.54898		0.582313		0.629252		0.62449		0.659864	
10	0.726531		0.741497		0.726531		0.77415		0.77483	
Mean	0.6557		0.6393		0.6982		0.6706		0.7204	
Standard Deviation	0.0781		0.0643		0.0654		0.0649		0.0479	

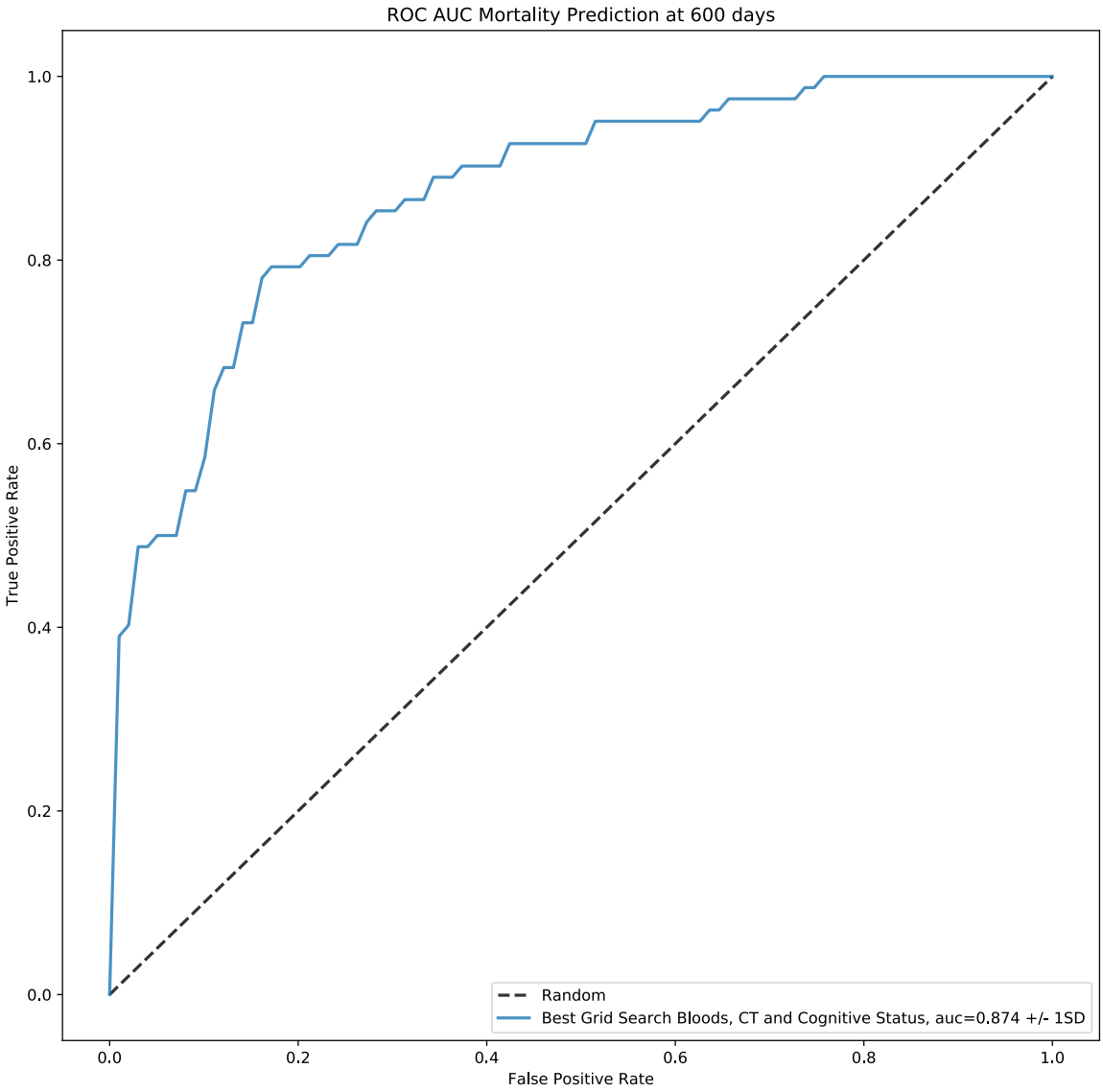
CV fold	CT intracranial, cognitive status, admissions, age & sex		CT extracranial, cognitive status, admissions, age & sex		CT whole brain, age & sex		CT whole brain, cognitive status, admissions, age & sex		CT whole brain, bloods, admissions, age & sex	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test
1	0.738095		0.834656		0.81746		0.829365		0.818783	
2	0.75		0.787037		0.800265		0.78836		0.825397	
3	0.900794		0.862434		0.89418		0.871693		0.862434	
4	0.815278		0.806944		0.830556		0.809722		0.826389	
5	0.773611	0.820164892	0.85	0.847732738	0.856944	0.860958434	0.829167	0.854087942	0.843056	0.87427001
6	0.872222		0.847222		0.883333		0.895833		0.888889	
7	0.880272		0.861224		0.914286		0.922449		0.912925	
8	0.734694		0.811565		0.819048		0.82585		0.789116	
9	0.77415		0.795918		0.764626		0.761905		0.767347	
10	0.834694		0.834014		0.855782		0.851701		0.881633	
Mean	0.8074		0.8291		0.8436		0.8386		0.8416	
Standard Deviation	0.0621		0.0271		0.046		0.0485		0.0455	

Fig 6.3: Rainplot representing hierarchical increasing predictive performance accuracy with increasing modalities



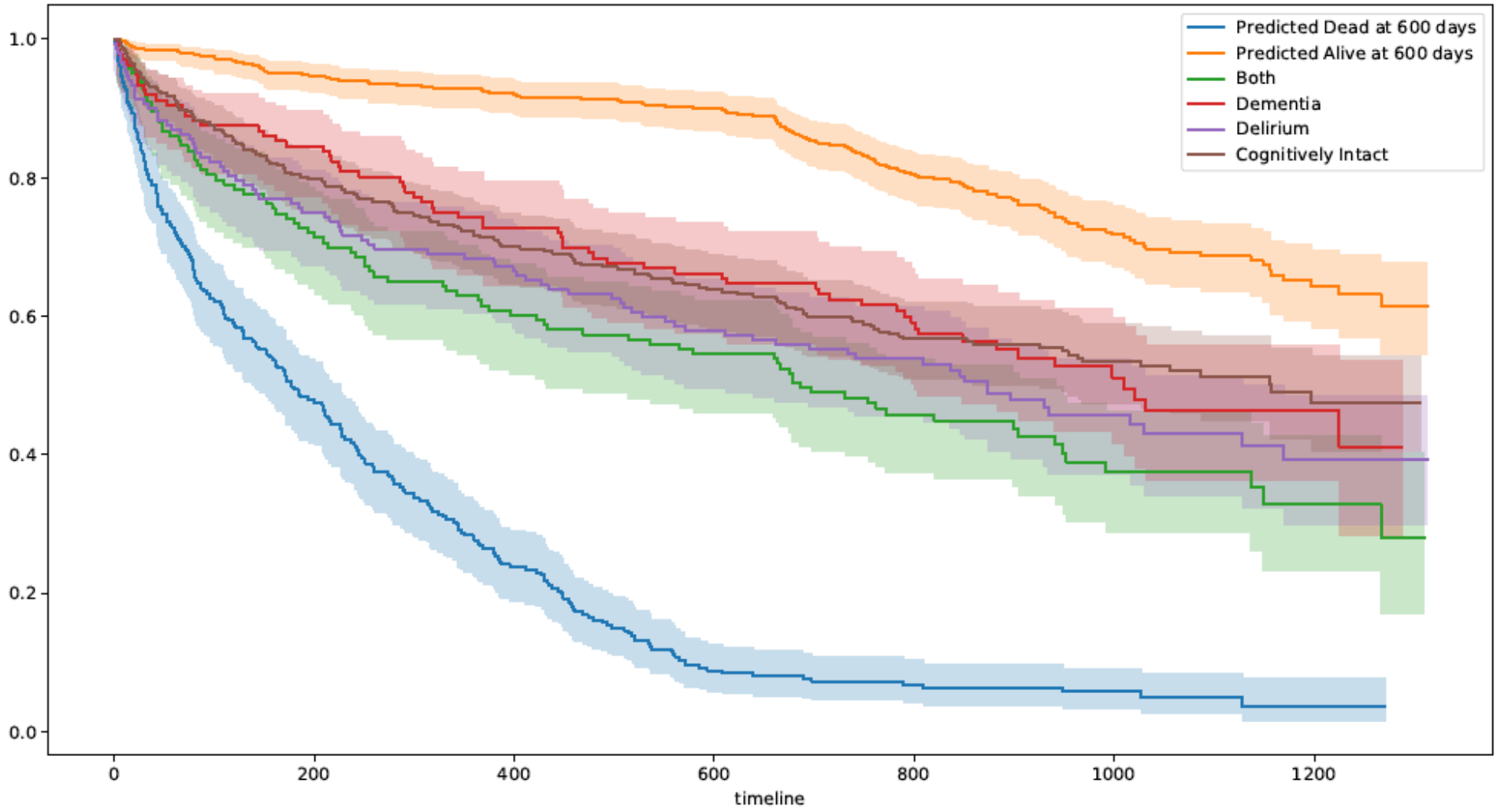
Red denotes AUCROC on held out test partition, intervals generated from ten cross validation folds for each model

Fig 6.4: ROC AUC of predicted mortality at 600 days



Area under the curve plot for best performing model (0.874); random chance denoted by dotted line

Fig 6.5: Re-application of prediction model to original dataset with Kaplan Meier of patients predicted alive (orange) and dead (blue), demonstrating superior performance compared to utilising stratification into four classes of acute cognitive status



6.3.3 Anatomical correlates of mortality

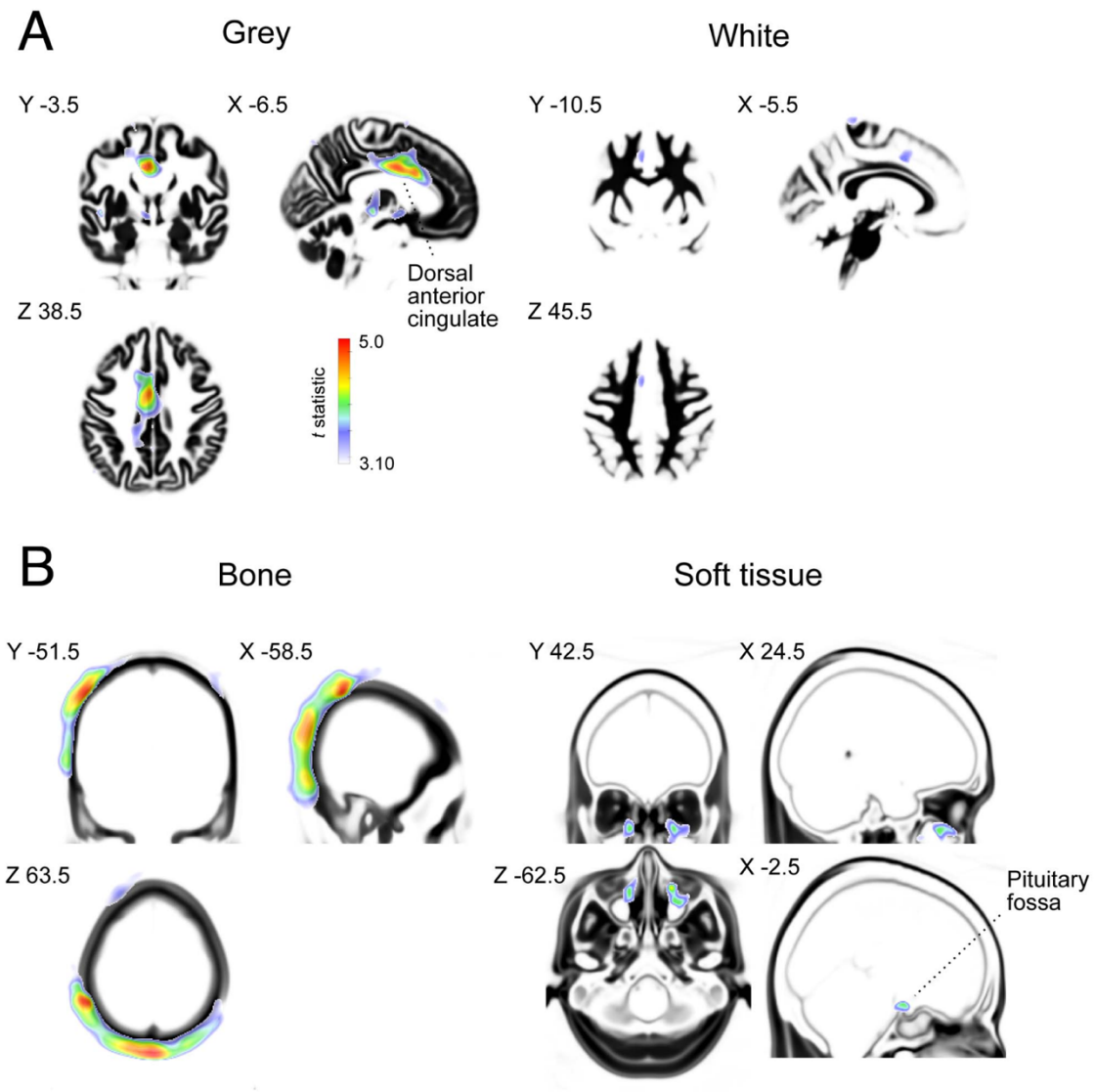
Sensitivity analyses confirmed the feasibility of my analytic approach, despite the biological heterogeneity of an unselected clinical dataset and the use of lower resolution CT instead of MR imaging. Comparing patients with and without dementia, there was significant loss of medial temporal grey and white matter in patients with diagnosed dementia, consistent with known changes in Alzheimer's disease, the most common dementia subtype,

Voxel-wise mass univariate models of ground-truth mortality revealed multiple loci of linear association distributed across the intracranial and extracranial compartments (fig 6.6). Intracranially, the most prominent associations were seen in the dorsal anterior cingulate grey matter. Extracranially, widespread differences was seen in the vicinity of the parietal and occipital bones, with further loci not surviving conservative multiple comparisons correction observed within the sinuses, and the region of the pituitary fossa.

6.3.4 Anatomical features of predictive importance

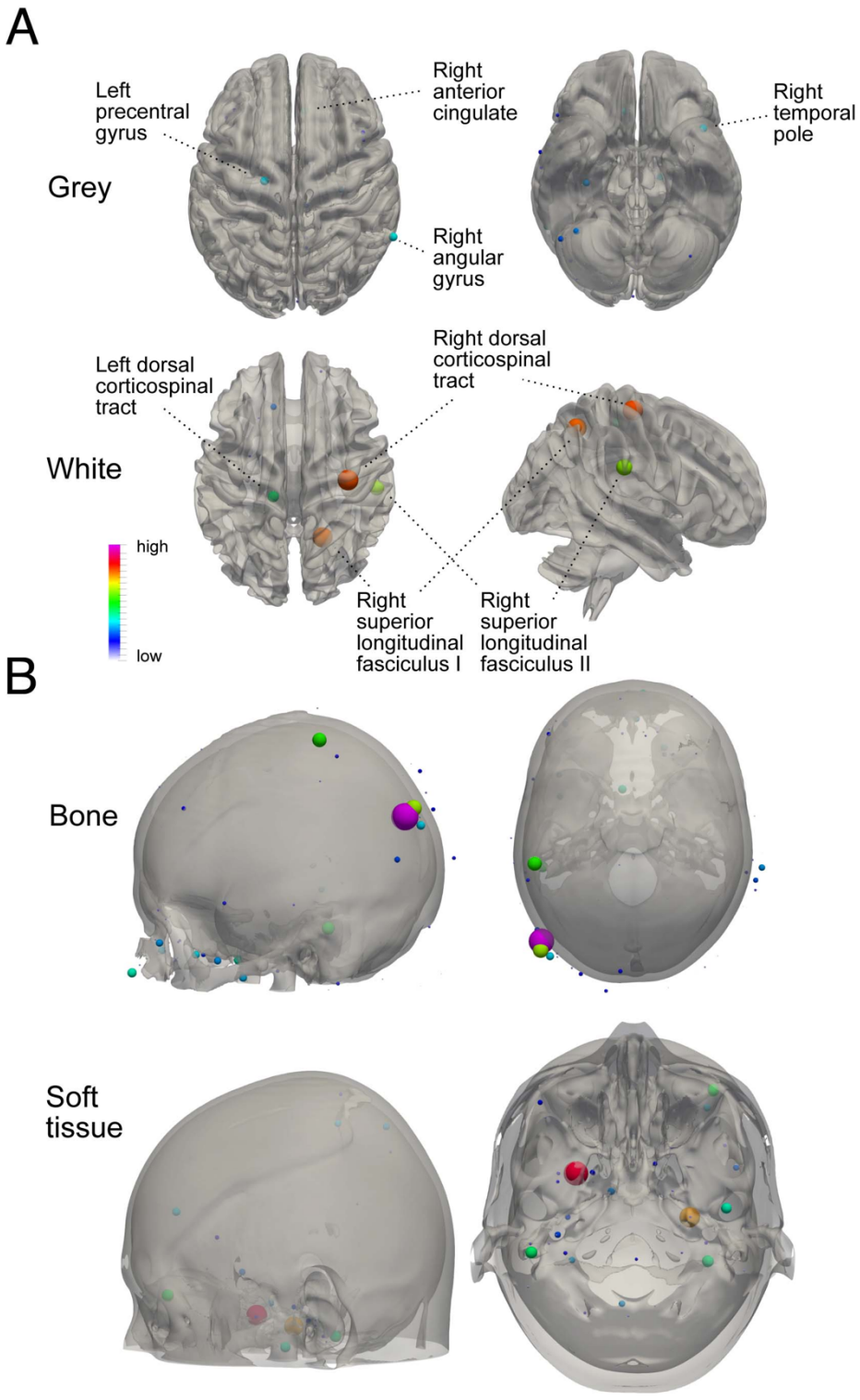
Projection of the anatomical feature importances derived from the best performing XGB model showed a widely distributed pattern of dependence (fig 6.7). Intracranially, left precentral gyrus, right anterior cingulate, right angular gyrus and right temporal pole, the region of the dorsal corticospinal tract, and the right superior longitudinal fasciculus were identified. Extracranially, parietal bone was highlighted as in the linear models and diffusely distributed soft tissue. However, I note that laterality of backprojected features should be interpreted with caution: the aim of high dimensional models is to select the best features for optimising task performance. In the case of bilateral features that contribute equally well, one would be discarded as redundant. As a result, it is not possible to ascertain whether a lateralising feature identified on high dimensional discriminative models is truly unilateral or represent a symmetrical bilateral contributor. Non-anatomical features of predictive importance are presented in table 6.6

Fig 6.6: Statistical parametric map (SPM) demonstrating voxels of significance comparing contrasts of ground truth alive vs dead status at 600 days



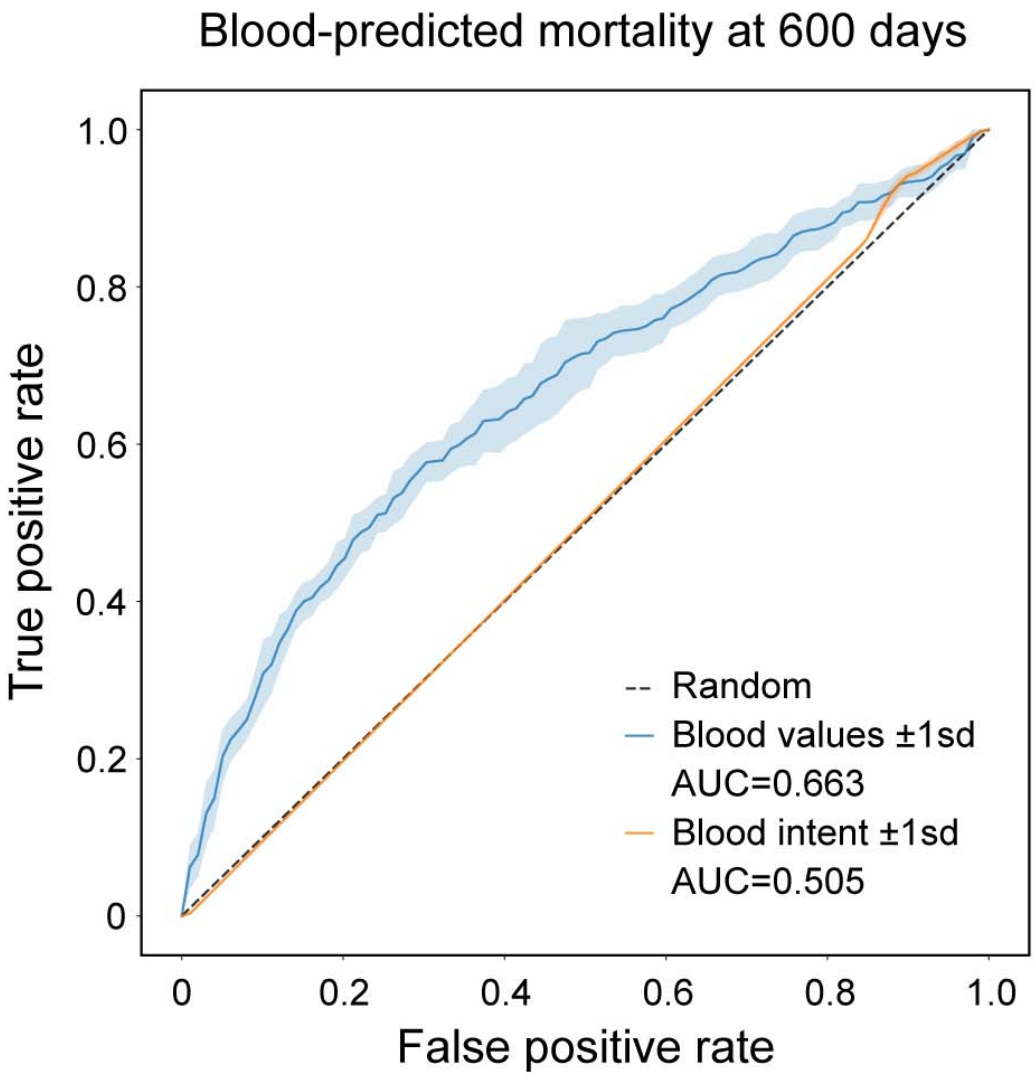
SPM demonstrating significant intracranial voxels (panel A) and extracranial voxels (panel B) on full adjustment by age, sex, dementia, delirium and total brain volume. T statistics of voxels demonstrated surviving full adjustment, with dorsal anterior cingulate and extracranial soft tissue identified as anatomical correlates of mortality.

Fig 6.7: Back-projection of feature importance of mortality



Most contributory intracranial (panel A) and extracranial features (panel B) to XGB predictive model back-projected onto brain template, with degree of importance denoted by colour spectrum

Fig 6.8: AUCROC of predicted mortality at 600 days for baseline model with summarised blood variables compared to using intention to investigate for bloods alone.



Summarised blood values demonstrating significantly greater contribution to prediction model than the intention to investigation of performing bloods alone, which is minimally more successful than random chance alone

Haematological and biochemical correlates of mortality

Albumin, urea, alkaline phosphatase, C-reactive protein, haematocrit and platelet count were the greatest contributors to the predictive model. The first albumin reading of a patient's admission was the most informative biochemical correlate, over fourfold more than the standard deviation of albumin and mean urea. Sensitivity analyses demonstrated the contributions of bloods over and above an intention to investigate effect, demonstrated by the improved performance of bloods only models when using summarised values instead of whether an investigation took place or not (fig 6.8)

Table 6.6: Feature importance of haematological and biochemical contributions to the predictive model.

Non-anatomical Feature	Predictive Model Importance	Importance ranking of model contribution
Albumin (baseline)	0.011143	15
Albumin (SD)	0.002525	108
Urea (mean)	0.002282	116
ALP (SD)	0.001650	148
Age	0.001122	180
CRP (baseline)	0.000799	219
ALP (mean)	0.000727	231
CRP (mean)	0.000567	259
Albumin (mean)	0.000220	350
Haematocrit (mean)	0.000168	374
Platelet (baseline)	0.000140	390

Predictive model importance of non-anatomical features and rank of feature importance to whole model including full neuroimaging features; CRP – C-reactive protein, ALP – alkaline phosphate, SD – standard deviation

6.4 Discussion

In this study, I examined a large, unselected cohort of acutely hospitalised older patients evaluated for the presence of acute or chronic cognitive impairment and imaged with cranial CT, then characterised the distribution of multimodal biological factors predictive of mortality over the subsequent 600 days. My findings quantify the predictive potential of cumulative multi-modal feature inputs for long term prognostication, while producing a clinical tool for medium-term mortality prognostication, operational at the point of admission and validated for the oldest-old patients.

6.4.1 The predictability of mortality from routine clinical data

Employing strictly out-of-sample evaluation of performance, I showed that 600-day mortality is predictable with high fidelity—AUCROC 0.874 —from the combination of basic clinical data with routine investigations. This suggests the presence of stronger predictive signals than are harnessed by current mortality prediction models—ranging from 0.51 (Di Bari, Balzi et al. 2010) to 0.78 (Curtin, Dahly et al. 2019)—especially when applied to patients aged 75 and over (Yourman, Lee et al. 2012, Curtin, Dahly et al. 2019). Of note, the signal here is grounded primarily in fundamental, quantitative biological characteristics of the patient rather than the circumstances of their care—such as mode and specialty of admission—minimising the risk of poor generalisability across other healthcare systems observed for higher-performing scores such as Hospital One year Mortality Risk score (van Walraven 2014) (van Walraven, McAlister et al. 2015) (Curtin, Dahly et al. 2019).

Our input features are not dependent on, or affected by, local healthcare trends, such as thresholds for emergency department admission, nor impacted by cultural norms such as the incidence of nursing homes admission for functionally impaired older patients. In contrast, alternative scores such as Hospital One year Mortality Risk score (van Walraven 2014) include population-specific features in its inputs, such admission by ambulance, number of emergency department visits and specialty of admitting service in its inputs, limiting its generalisability: very high levels of discrimination were initially reported in derivation (C statistic 0.92) (van Walraven 2014) and internal validation cohorts (C statistic 0.89-0.92) (van Walraven, McAlister et al. 2015), but significantly attenuated on independent validation in an Irish cohort, with C statistic of 0.78 (Curtin, Dahly et al. 2019).

The limited contribution of age alone to mortality prediction in our model is in keeping with previous prognostic models (Teno, Harrell et al. 2000). In short term prediction models such as

APACHE-III, age only made up 3% of the score input despite being derived from intensive care unit patients with a wide age range (Knaus, Wagner et al. 1991). There is increasing recognition of the benefits of including care needs and physical function to mortality prediction models. HOMR produces a composite score of Charlson comorbidity index with age (van Walraven 2014), while HOMR-NOW further includes additional measures such as rehabilitation or nursing home admission (van Walraven and Forster 2017). The Health Assessment Tool, frailty index and walking speed are among the most accurate predictors of mortality in older patients (Zucchelli, Vetrano et al. 2019). However, subjective, time-consuming assessments and the wide variety of factors that lead to functional impairment, limit the use of these instruments. Objective measures of underlying aetiologies that lead to functional deficits, for example neuroimaging that may reflect volumetric loss associated with impaired cognitive functions, may offer a compromise. This would explain the high performance of this study's prediction model. A critical contribution came from intra- and extracranial characteristics captured by CT imaging—a quantitative modality—that are reproducibly identifiable across clinical environments. Robustness to clinical and demographic variation naturally requires evaluation in other population samples, and I plan to validate our findings in an independent cohort formally. However, the level of observed fidelity motivates further exploration of such predictive models which can be implemented within the existing clinical pathway without disruption to established care.

Simple predictive models are constitutionally incapable of integrating information distributed across multiple interacting factors. Where, as here, the causal field is plausibly wide and densely interdependent, statistical models of greater dimensionality and flexibility are required. Exploiting the flexibility and robustness of gradient-boosting machines, I have shown that escalating model dimensionality is rewarded by more accurate predictive performance—quantified out-of-sample. This indicates the presence of distributed, possibly interacting, factors that collectively strongly predict mortality even if they may be only weakly predictive in isolation. The scale of the informative dimensionality—thousands of variables—suggests room for improvement with more data and finer model architectural tuning. Note that any substantive increment in predictive fidelity is valuable for a model applied at the individual patient level—an imperfect model here can never be too accurate. The greater sensitivity of complex models to distributional shift is no longer insuperable: it is addressable algorithmically, and through expanding the scale and diversity of modelled data.

6.4.2 Possible mechanisms of increased mortality

Mass-univariate analyses of haematological, biochemical, and especially imaging features implicate a unifying association of mortality with sarco- and osteopenia, as seen through modulation of parietal and occipital bone and cranial soft tissue involvement. Sarcopenia, defined as the progressive generalised loss of skeletal muscle, is associated with increased mortality risk (Cruz-Jentoft, Bahat et al. 2019), potentially mediated via impaired mobility, falls and respiratory complications (De Buyser, Petrovic et al. 2016, Bone, Hepgul et al. 2017). Similarly, osteoporosis and the increased fractures are well recognised to be associated with increased mortality risk: neck of femur fractures are associated with one-year mortality of around 30% (Roberts and Goldacre 2003). Albumin is commonly acknowledged to be a poor marker of nutritional status, particularly in the acute setting. In this context, albumin plausibly represents altered hepatic synthesis in favour of acute phase proteins indicating proinflammatory acute illness. The contributions of urea and alkaline phosphatase are consistent with possible links to sarcopenia and bone density respectively.

Within the brain, striking involvement of medial frontal cortical areas implicated in voluntary motor behaviour and autonomic function principally the anterior cingulate, may be explained by the potential impact of dysfunction in either domain.

It is reasonable to assume that in the absence of an acute ischaemic or traumatic intracerebral event, CT imaging is informative about baseline cognitive vulnerability but not the acute physiological insult. How highlighted brain regions contribute to vulnerability to medium-term mortality risk is not necessarily straightforward. It may reflect the importance of specific cognitive functions from particular brain regions for medium-term survival. Cognitive deficits from multiple brain regions would present as impairments in physical function, imprecisely quantified by poorer performance in functional indices such as the Barthel Index. Either alternatively or concurrently, it may be that we are observing the reduced capacity of a brain region to compensate for damage or atrophy. Sparing of frontal lobes may not necessarily suggest that frontal lobe function is less critical to survival but rather reflects greater redundancy to mitigate volumetric loss.

The importance of motor function is reinforced by the enhanced multivariate feature importance of voxels falling within the primary motor cortex and the corticospinal tract. These changes may be explained not only by brain pathology, such as small vessel disease that either directly or indirectly reduces grey matter concentrations locally, but potentially by neural adaptation to long term immobility of non-neural origin. The involvement of areas with dense (angular gyrus) or remote

(temporal pole) connectivity suggests modulation here may also reflect differential rates of length-dependent degeneration in white matter and the grey matter it connects.

The implication of motor and somatosensory cortices may be hypothesised to result in motor weakness and impaired sensory perception, increasing the risk of falls and dangerous consequences such as fractures resulting in death. Neck of femur fractures are associated with one-year mortality of around 30% (Roberts and Goldacre 2003). Significant differences in medial temporal lobe volumes between prediction groups are consistent with typical patterns observed in Alzheimer's dementia (Jack, Petersen et al. 1998), as is the specific implication of the left caudate nucleus, which is in keeping with asymmetric volume loss observed from MRIs of patients with dementia (Barber, McKeith et al. 2002, Jiji, Smitha et al. 2013). The left caudate is important in language monitoring and control (Crinion, Turner et al. 2006), while bilateral insula volume loss has is associated with non-fluent aphasia dementia syndromes (Mandelli, Vitali et al. 2016). This pattern of regional differences suggests particularly important mechanistic contributions of impaired short-term memory, working memory and language towards medium-term survival in older patients. These inferences may appear at odds with the lack of improved model performance with addition of acute cognitive status. However, this may reflect either the poor diagnostic rates of dementia, or the contributions of subclinical cognitive impairments to mortality, which while not overtly presenting as dementia syndrome, results in other subtle cognitive and consequently functional deficits impacting mortality risk. Further exploration of these biologically intelligible patterns is merited.

6.4.3 A multimodal index of frailty

A predictive model with high fidelity at the individual level has the potential to support clinical decision-making. Here quantifying the risk of death at the outset of a hospital admission enables proportionate pre-emptive action—by both patients and clinicians—to minimise it. Grounding an index of frailty in multimodal signals in principle renders it more robust to incidental variations by broadening its evidential support, and potentially widens the field of manipulable factors critical in any one patient, permitting more closely individuated interventions. Hardening the predictive model to missingness and distributional shift, and incorporating machinery for causal inference, require further algorithmic development with larger-scale data that this proof-of-concept now justifies. Crucially, since routine clinical and investigational data models appear to be sufficiently powerful here, real-world implementation of a decision-support tool does not require any changes to clinical pathways, substantially lowering barriers to implementation. Indeed, it may be argued

that clinicians have a moral duty to maximise the guiding intelligence extracted from the data they obtain from patients, often at individual, and always at institutional, cost.

A high-fidelity individual index of frailty also has applications in stratifying patients in observational and interventional research studies, where unmodelled structured variability could otherwise conceal or distort inferred effects. Furthermore, multimodal models may reveal heterogeneities between subpopulations exhibiting the same risk, suggesting potential differences in causation that would confound inferences unless explicitly modelled. Attention to heterogeneity is paramount in the older patient, where the multiplicity and diversity of observed pathologies is high.

6.4.4 Strengths and limitations

The synthesis of multimodal signals spanning demographics, clinical features, blood tests, and CT imaging data is unique among prognostic models in this population. The use of objectively quantifiable features derived only from routinely collected data without the need for pre-admission information or potentially subjective clinical assessment is a central strength, promoting generalisation across healthcare systems and enabling implementation without disrupting established pathways. In addition, this cohort captured the acute admissions behaviour of oldest-old patients, with a mean age of 85.5 years. Model development and out-of-sample validation on one of the largest unselected cohorts of older patients evaluated for acute illness in frail patients is grounds for confidence in the robustness of the findings. Participants were observed for at least 20 months prior to determining the primary endpoint (mortality) to study medium-term mortality risk. This resulted in roughly equal group sizes of alive or dead individuals at the point of censoring.

An array of limitations should be noted. In keeping with all observational analyses of routine clinical data, a degree of corruption by (potentially structured) missingness, acquisition and documentation errors, and clinical uncertainty is inevitable. Minimal improvements with addition of acute cognitive status was likely caused by inaccurate delirium and dementia diagnoses. Formal consultant delirium diagnosis did not guarantee the use of a validated diagnostic tool: the use of a delirium diagnostic instrument was not compulsory and may have contributed to inconsistency among diagnoses between different clinicians. Meanwhile, dementia under-diagnosis may have resulted from insufficient symptom recognition, use of corroborative imaging and healthcare engagement secondary to a variety of clinical and social factors. While medical records and collateral histories may be sufficient for a number of patients, this was not necessarily confirmed with a validation process such as IQCODE or formal corroboration with primary care records. Inherently, both required prior interaction with healthcare services – a lack of dementia diagnosis

secondary to lack of contact with healthcare practitioners did not equate to zero probability of clinical dementia.

Equally, though the cohort is fully-inclusive of the clinical stream, only those with CT imaging of the head, carried out for indications individual to each patient, were retained in the analysis. We pursued this approach to maximise ecological validity, replicating the quality of data a real-world institution would naturally see. The impact of potential biases is minimised by the use of sequential, unselected data, enabling inference across all those in receipt of the criteria investigations. I explicitly quantify the effects of intention-to-investigate in relation to individual blood tests, finding it to contain negligible predictive signal. Institutional-level variability in clinical practices could impact generalisation and need exploration in future multi-centre studies.

Third, the study offers only a view of relationships between a cross-sectional snapshot of biochemistry and plain CT structural neuroimaging with medium-term mortality risk. Longitudinal changes in biochemistry, neuroimaging and cognitive statuses with later-life mortality risk could not be investigated with this dataset.

While the alternative use of structural or functional MRI instead of CT may have increased dimensionality and feature complexity, potentially improving model performance, finite resource allocation would limit the feasibility of this in practice. In keeping with all observational analyses of routine clinical data, a degree of corruption by (potentially structured) missingness, acquisition and documentation errors, and clinical uncertainty is inevitable. Finally, because it was not available, we do not model the cause of death, only the primary diagnosis on the admission that triggered entry into the cohort. There is no reason to expect the distribution of causes of death to differ substantially from that observed for the underlying population, and my focus here is the fact of it rather than its cause.

6.4.5 Conclusion

Examining the distribution of a wide array of multimodal predictive factors in older patients acutely admitted to hospital and imaged with cranial CT, I have demonstrated proof-of-concept of utilizing a high-dimensional approach to predictions in older patients and established the foundations of a multimodal approach to predict long-term mortality, operational at the point of admission and validated for unselected oldest-old patients. Our analysis demonstrates the benefit of using machine learning to enable models incorporating multiple modalities, highlighting the predictive potential of defining “multimodal frailty”. Inferences drawn from the prediction model suggest the importance of unimpaired higher motor and autonomic function towards medium-long term

survival. Taken together, this proof-of-concept study demonstrates the degree of performance accuracy possible by using advanced machine learning techniques on high-dimensional input features to predict mortality as well as make mechanistic inferences on the basis of extracranial and intracranial contributions of radiological features to these models.

7 Precise Identification of delirium Subtypes through Clinical Analyses (PISCA)

Chapter Outline

- Rationale for clustering
- Methods
  - Data projection to 2-dimensions
  - Cluster selection
  - Model prediction
- Results
  - Hierarchical cluster selection
  - Comparing survival and length of stay predictions by clusters and full features
- Discussion
  - Structure of acute decompensation in older patients
  - Importance of clustering
  - Strengths and limitations of PISCA

Chapter 6 demonstrated the potential of high-dimensional, multi-modal machine learning to predict long term outcomes in older people presenting with acute illness. While state of the art, this approach is inherently computationally intensive and such resources may not be available in all healthcare settings. Clinically useful but albeit less impressive performanc, may be possible using a condensed set of features, requiring significantly less intensive computation. In addiiton, reduction into broader stratified subtypes may be sufficient for purposes such as clinical trials. This chapter is a preliminary study that explores the possibility of clustering as a middle ground solution. I will first describe the rationale for clustering.

7.1 Rationale for clustering

The delirium syndrome is inherently complex in its pathophysiology and heterogeneous in its clinical presentation. Despite this, it is commonly only operationalised as a dichotomous diagnosis, as the presence or absence of delirium. A binary diagnosis of delirium offers clinically useful information for mortality risk, when ascertained as cross-sectional incidence (Witlox, Eurelings et al. 2010) or period prevalence of delirium (Tsui, Searle et al. 2022). Yet there is likely further granularity within the arousal, attentional, cognitive, motor, neuropsychiatric and functional deficits

in delirium. Longitudinally, as demonstrated in earlier chapters, baseline features such as premorbid cognition and measures of frailty prior to acute illness also significantly contribute to the presentation and sequelae of delirium.

While utilising the full spectrum of detailed cross-sectional and longitudinal measures in an acutely unwell patient may result in more accurate predictions of adverse outcomes (Chapter 6), there are further ways of harnessing the breadth and detail of high dimensional data. As described in Chapter 4, unsupervised machine learning techniques offer data-driven approaches to cluster data, offering a novel method to define clinical subtypes agnostic to clinical precedence.

However, if greater numbers of input features can be anticipated to produce better-performing predictive clinical models, what are the additional advantages of clustering and reducing data dimensionality? First, while higher number of dimensions improve model performance, this is true up to an optimal point, after which further dimensions add more noise than signal to a model, and consequently model performance diminishes. Second, clustering may produce relatively accurate predictive performance, albeit less accurate than high dimensional and complex algorithms, yet with significantly lower computational requirements and hence, easier to implement in settings with limited resources. Third, clustering into clinical subtypes allows division into broad groups of treatment strategies. Fourth, subtyping may lead to greater pathophysiological understanding: do different delirium subtypes result from different aetiologies, or is delirium defined by a common final pathophysiological pathway? Fifth, clustering allows accurate stratification for future clinical trials: do different drugs have differential effects on patients with different delirium subtypes?

Lastly, clustering would challenge our current definition of delirium. The current DSM syndrome is a consensus definition difficult to operationalise, can be subjectively interpreted and is imperfect in relation to predicting adverse outcomes. For example, there is no consensus on how best to objectively measure " disorganised thinking”. In contrast, high dimensional clustering can instead select the most important, relatively objective features, guided by their contribution towards an adverse outcome. This could lead to a more empirically-driven, clinically useful term for use in acute illness, beyond the current accepted definition of delirium.

The most common approach to an unwell patient with an acute confusional state is to divide delirium by motor phenomenology, into hypoactive, hyperactive and mixed delirium, as initially proposed by Lipowski (Lipowski 1983). These subtypes have been reported to demonstrate clinical utility: hypoactive delirium across an older population was associated with higher mortality risk, with altered arousal associated with 6-fold mortality risk (Todd, Blackley et al. 2017) (Jackson, Wilson et al. 2016). In intensive care settings, hypoactive delirium was associated with worse

cognitive decline (Hayhurst, Marra et al. 2020). However, cross-sectional subtyping by motor phenomenology alone poses a number of limitations. First, associations with adverse outcomes are potentially heavily confounded by factors that may contribute to reduced motor activity, for example, prescription of sedating medications, particularly in enhanced care settings. Hypoactivity may also be confounded by greater illness severity and increased likelihood of a palliative setting (Meagher, Leonard et al. 2011).

Secondly, except for defining an “acute change” in cognition, premorbid baselines of cognition or function are not currently routinely utilised. Chapter 5 demonstrated that a patient’s cognitive and functional baselines are essential in contextualising acute confusion to predict adverse outcomes optimally. As demonstrated in earlier analyses, the effects of delirium on cognitive impairment and mortality extend beyond a linear dose-dependent relationship. In addition, the contributions of function and social environment on cognitive outcomes are increasingly well-recognised. the modifying effects of frailty on the neuropathology–cognition relationship (Wallace, Theou et al. 2019).

Third, using motor phenomenology emphasises clinical attention on motor symptoms, instead of the nature and severity of an attentional or arousal deficit syndrome, which is pathognomic of delirium. Similarly, some patients may not necessarily satisfy diagnostic criteria of delirium, but certainly present with features of an acute confusional state with attentional deficits in the context of an acute precipitant. Commonly referred to as subsyndromal delirium, these patients would otherwise be excluded from clinical prognostication despite likely demonstrating features with some degree of clinical utility.

Last, the delirium syndrome is currently routinely described using clinical features alone. Other cross-sectional modalities such as blood biomarkers, better articulating acute illness severity and premorbid frailty are not current features of the DSM-IV definition.

7.2 Methods

In this chapter, I will use the clinical data from the Delphic study, described in Chapter 5, to identify clusters and compare their clinical utility against current standards. The dataset was a combination of binary (e.g. sex), ordinal (e.g. MDAS item score) and continuous (e.g. blood results) features from cross-sectional assessments during acute illness and previously ascertained baseline parameters. I first projected this multidimensional dataset onto a two-dimensional manifold for ease of cluster visualisation. Similar data points were then designated as homogenous clusters using hierarchical clustering. Third, I constructed increasingly complex models to predict survival

and length of stay, comparing performances between baseline model of age, sex and delirium status, replacement of delirium status with clusters, and finally, full inclusion of individual features. Last, I interrogated the components of each cluster to infer mechanistic insights.

7.2.1 Data pre-processing

As for the main DELPHIC analyses, inpatient assessments were grouped by admission IDs. Features with more than one data collection point per admission were summarised into first, maximum, mean and standard deviation derived variables. Missing data was median imputed. The entire dataset was standardised.

7.2.2 Autocorrelation

Correlation coefficients were calculated between each clinical feature pair. These were then visually checked for autocorrelation in a plotted heatmaps.

7.2.3 Two dimensional manifold

First, I transformed the multidimensional dataset into a two-dimensional manifold. To reduce dimensionality to allow efficient t-distributed stochastic neighbour embedding (TSNE), the full set of derived clinical features were downsampled to the first 50 principal components, a compromise value between sufficient explanation of variance and computation requirements. I performed TSNE with a range of perplexities (5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100. Individual two-dimensional manifolds plotted for visual inspection. Lastly, after perplexity = 20 was selected as the most appropriate TSNE setting. Values for each clinical feature for each participant was plotted onto the two-dimensional manifold, to inspect if clustering was inherently over-contributed to by individual features.

7.2.4 Cluster selection

A dendrogram of agglomerative clustering was plotted using ward linkage and Euclidean affinity. The most optimal number of clusters to aim for was visually selected as the number of clusters transected by a horizontal line at the point of the dendrogram with the greatest vertical distances between clusters. The optimal cluster size was 10 and 12, with n=11 used to plot against the two-dimensional TSNE manifold.

7.2.5 Model prediction

Cluster designation was added to the pre-processed dataset. In order to accommodate uncertainty in the optimal number of clusters, predictions were performed using three to thirteen hierarchical clusters. Datasets were randomly split into training (70%) and testing (30%) partitions. The test partition contained unique patients only, with duplicate admissions for the same patient dropped, keeping on the first chronological episode. A predictive model using XGBoost was constructed for two outcomes: survival and length of stay. For survival, an accelerated failure time model was used instead of mortality at a timepoint, due to the imbalanced number of alive and dead participants by the end of the study. XGB regressor was used for length of stay.

Each model was trained and hyperparameters optimised using ten-fold cross-validation from the training partition. The best-performing fold hyperparameters on cross-validation were used to quantify performance on held-out test data. The evaluation metric used was negative log likelihood (nloglik) for survival and root mean squared error (RMSE) for length of stay.

For the survival model, lower and upper bounds were defined: if the patient was not censored by death, the lower bound was the study length, defined as the duration between date of baseline assessment and 1st June 2021, while the upper bound was the study length to infinity. However, if the patient was censored by death, the lower bound and upper bound were both the duration of the time the participant was in the study from baseline assessment until the date of death.

XGB hyperparameters were optimised using grid searches. The hyperparameters selected for the XGB accelerated failure time model were:

- Objective: 'survival:aft',
- Evaluation metric: 'aft-nloglik',
- Learning rate': 0.05,
- Max depth: 3,
- Subsample: 0.5,
- Minimum child weight: 5,
- Column sample by node :0.5

For all three outcomes, prediction outcomes using an incremental number of clusters, from three to thirteen, were compared against a baseline model including only age and sex, and a high-dimensional model including all features. The interquartile range and mean of the ten cross validation training folds and the test result were plotted onto a comparative rainplot.

7.2.6 Sensitivity Analyses

All steps described in 7.2.5 were repeated in a series of sensitivity analyses, varying how binary (e.g present or absence of a cardiovascular diagnosis), categorical or ordinal variables have been used. First, all data were treated as continuous, assuming that all binary features were ordinal. Next, binary and categorical features were instead treated as categorical data, resulting in mixed data types when incorporated with continuous variables. Subsequent clustering was performed using a Gower matrix. Last, binary clinical features were excluded, only ordinal categorical variables and continuous features were included, with all features treated as continuous values instead of one-hot encoding for categorical variables. This last approach demonstrated the best clustering and predictive outcomes: this was thus selected as the definitive approach.

7.3 DELPHIC clustering results

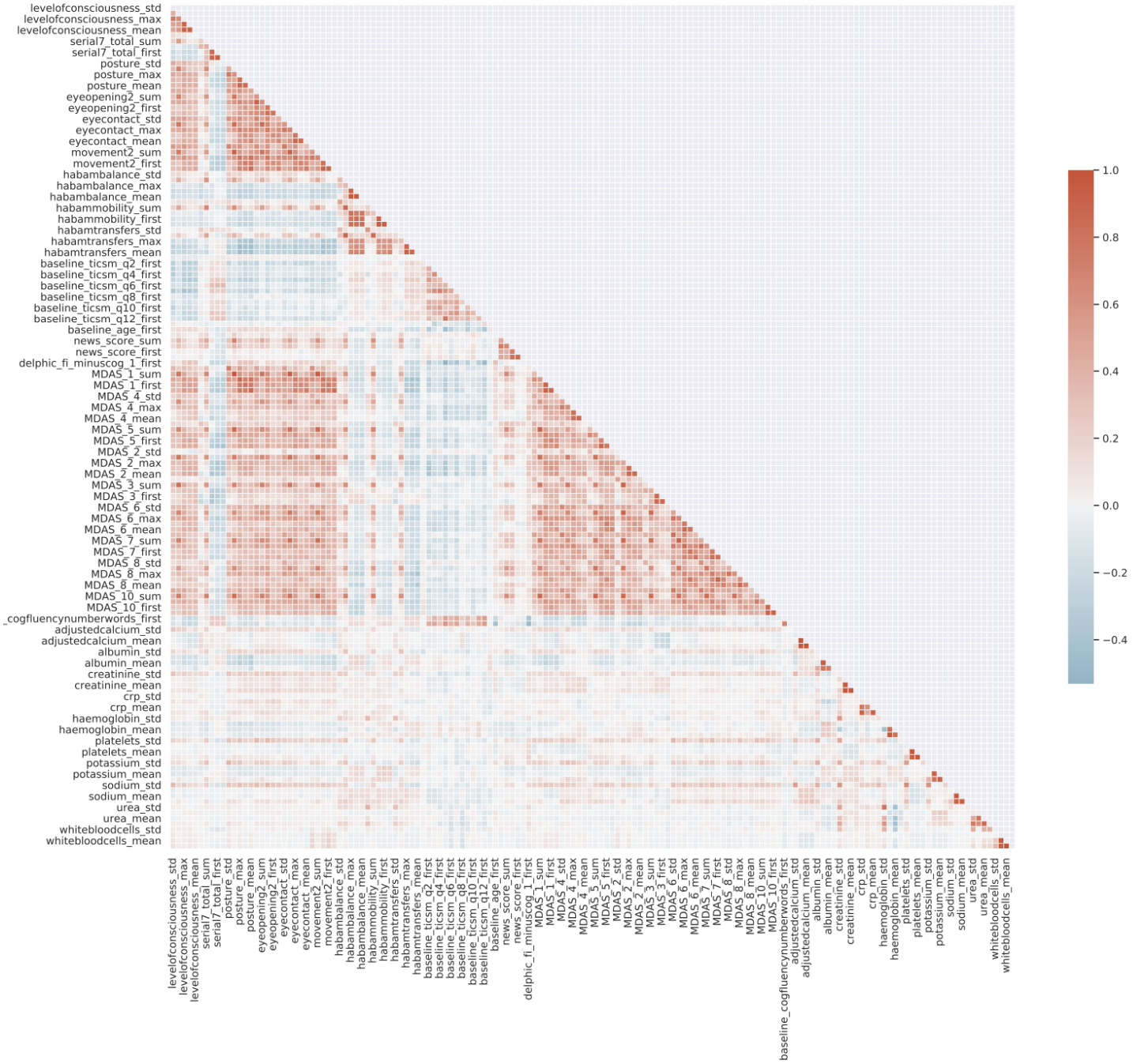
From the 1510 participants who underwent a baseline assessment, 209 were admitted during the course of the Delphic study for a total of 1999 episode days, over 371 unique admission episodes. The median time within the Delphic study was 860 days (IQR: 539 to 1141 days). By the end of the study, 73 deaths were recorded among participants who had been admitted to hospital at least once.

In total, 152 features were used for clustering, comprising 144 summarised variables originating from baseline ascertainment and inpatient assessments.

7.3.1 Autocorrelation

Included features were first inspected and sense-checked for autocorrelations. Autocorrelation was demonstrated between summarised derivations of the same variables (for example, the first, mean, standard deviation and maximum values of MDAS item 1) (fig 7.1). MDAS items positively correlated with OSLA scores, while both were negatively correlated to baseline TICS-M questions. Specifically, MDAS 1 (level of consciousness) was positively correlated to HABAM scores of transfers and mobility, consistent with gating of mobility by a minimum level of arousal. Minimal correlations were demonstrated between other clinical features.

Fig 7.1: Autocorrelation heatmap of features used for clustering

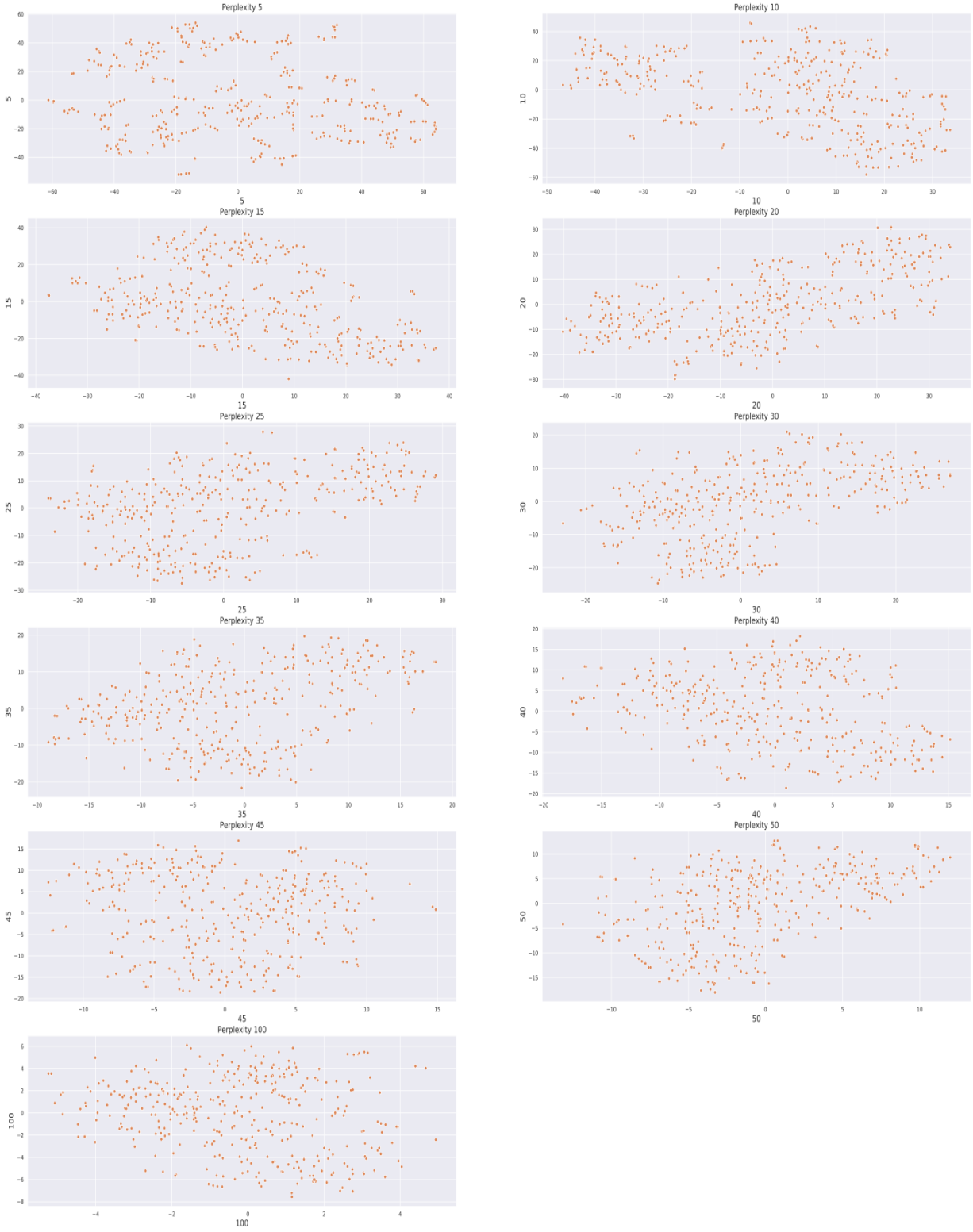


Colour bar indicative of correlation, with dark red indicating correlation coefficient of 1 (perfect positive correlation)

7.3.2 Perplexity Selection

After compression from multi-dimensional into a two-dimensional manifold, I selected a perplexity value of 20 on visual inspection for data transparency and sparseness, noting reasonable spatial distribution without obvious clumps of data. While there is no statistical method of determining the most optimal perplexity value, 20 is in keeping with between the range of 5 and 50 as recommended by TSNE authors (van der Maaten and Hinton 2008), and also consistent with $N^{0.5}$ = 19.26 as per convention (Oskolkov 2019).

Fig 7.2: TSNE with perplexity range from 5 to 100



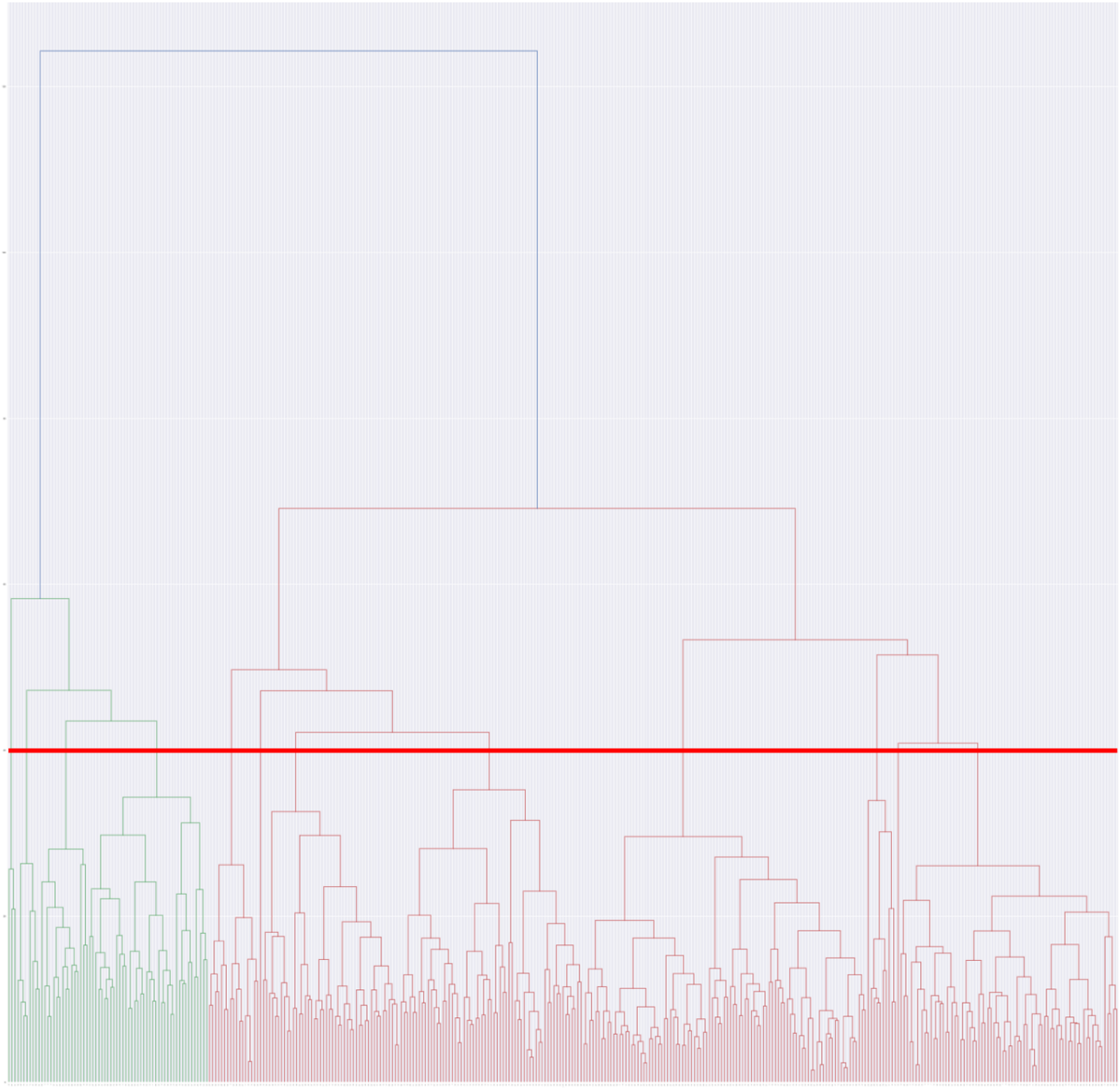
Visual inspection of local “clumps” and global transparency resulted in perplexity of 15-25 being chosen: with $n^{0.5}$ being almost 20, perplexity 20 was used for the final TSNE plot

Individual topographic clusters in 2-dimensional representations were not heavily influenced by information arising from either imputation status of a feature (supplementary fig 1) or contributions of a specific feature (supplementary fig 2). All items of the baseline TICS-M except item 5 (immediate recall) appeared to contribute similar topographical information. MDAS items 1 (level of consciousness) and 2 (disorientation) were similar in 2-dimensional representation, as were the baseline fluency items.

7.3.3 Cluster size

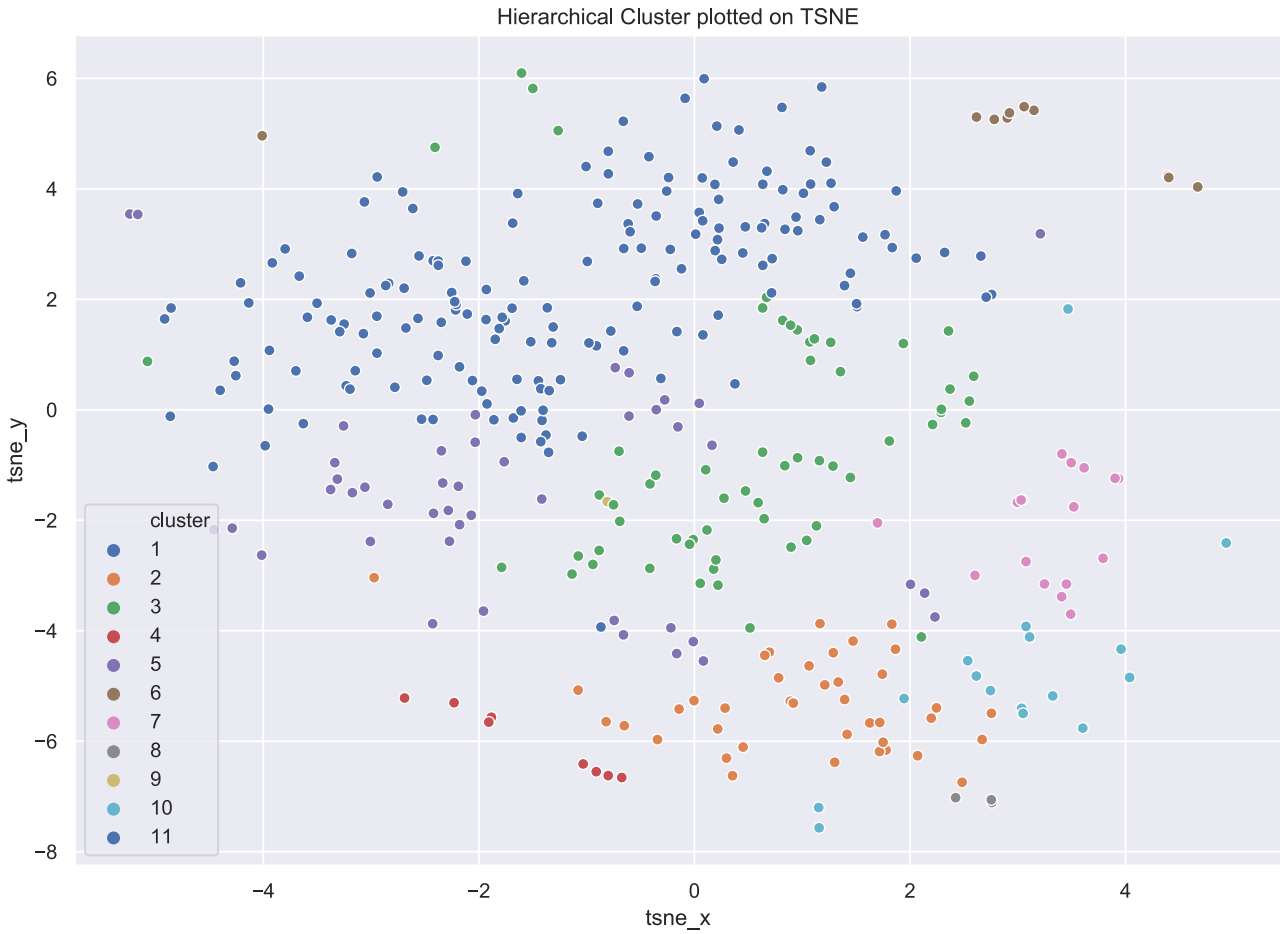
The mid-point of the maximal vertical distance on a plotted hierarchical dendrogram of agglomerative clustering crosses 10 to 12 clusters (fig 7.3). Subsequently, when eleven clusters were defined and replotted onto the two-dimensional TSNE manifold with perplexity 20, clusters with relative spatial distinction could be demonstrated in two-dimensional space (fig 7.4).

Fig 7.3: Agglomerative hierarchical cluster plot



Horizontal transection of longest vertical distance in plot (red line) cuts through approximately 10 to 12 clusters

Fig 7.4: Visualisation of dataset onto two-dimensional manifold by TSNE with individual data point marked by cluster definition

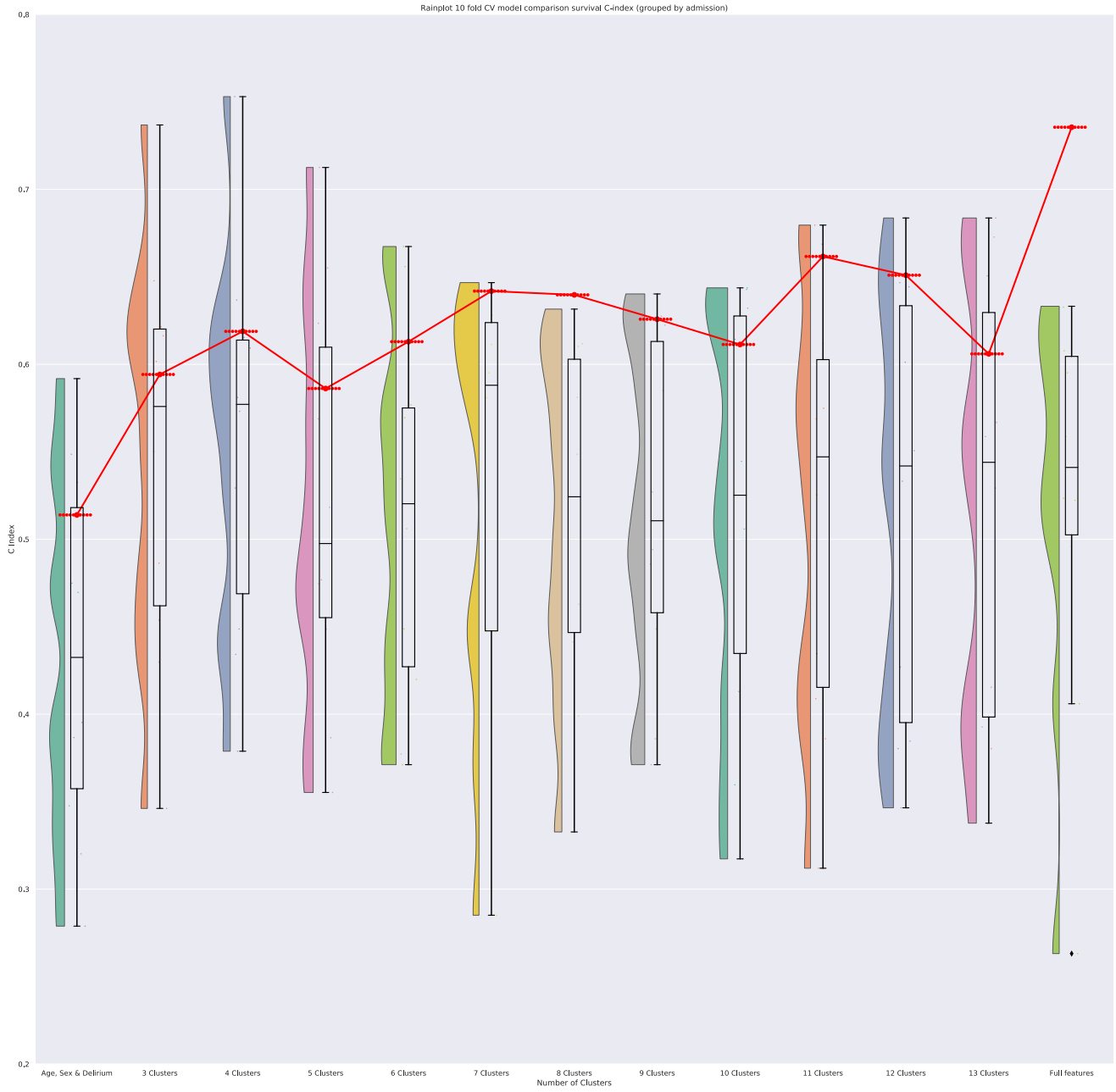


7.3.4 Survival prediction

A trend was demonstrated for increasing survival prediction performance from baseline models (age, sex, delirium status alone), addition of clusters instead of delirium status, and finally, all features.

Models including only age, sex and delirium status achieved a test C-index of 0.54. Addition of clusters, instead of delirium status, improved test performance, regardless of which of 3 to 13 clusters were used (fig 7.5). The model with the best test performance utilised 11 clusters, achieving a test C index of 0.66. The best survival prediction performance was achieved with the complete model, achieving test C-index of 0.71. Different trajectories of survival for participants of each cluster are demonstrated in figure 7.6.

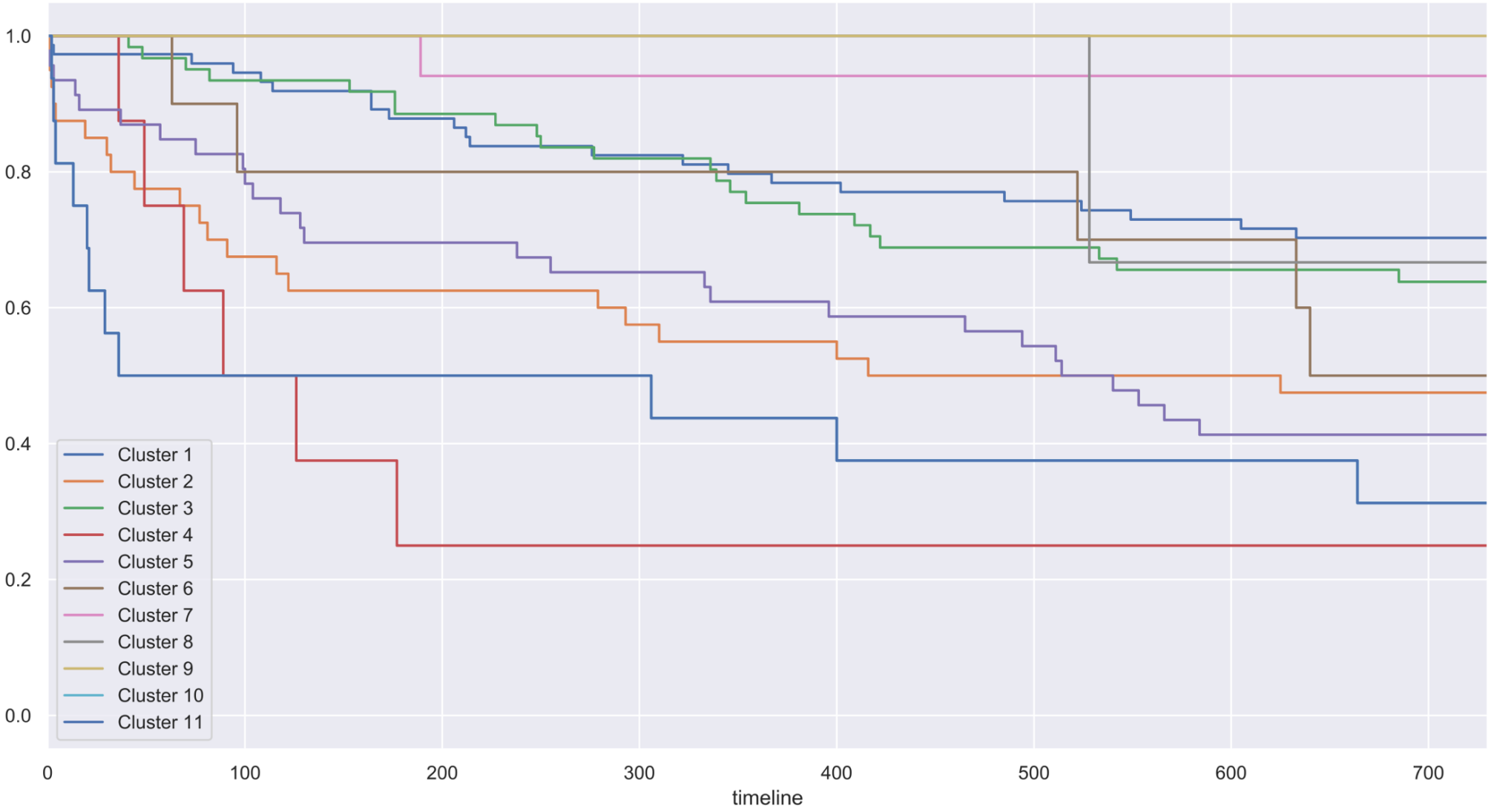
Fig 7.5: Rainplot comparing survival prediction (C-index) by baseline (age, sex, delirium status), with additional clustering and full features.



Rainplots and confidence intervals generated from ten-fold cross validation folds.

Red line denotes C-index on held out test partition

Fig 7.6: Kaplan Meier comparing by designated cluster



Further interrogation of the clusters identified the clinical contributors to each subtype (supplementary fig 3). I will highlight the four most populous clusters: Cluster 1, composed of 74 admission episodes from 55 unique participants, had participants with better baseline cognition and less frailty. Biochemically, they demonstrated higher albumin and lower creatinine, while also associated with higher CRP and platelet counts, suggestive of a pro-inflammatory picture. Their delirium severity and attentional deficits scores were low, with minimal working memory deficits. Participants in this cluster tended towards a relatively good survival outcome. Similarly, Cluster 3, (61 admission episodes from 33 unique participants), also demonstrated good survival outcomes. However, these participants had poor baseline cognition, high frailty, high levels of arousal attentional deficits and measures of delirium severity restricted to MDAS items 1 to 4 (measures of attention, arousal and disorganised thinking) but low scores for MDAS items 6 to 10 (neuropsychiatric and sleep-wake symptoms). Biochemically, they demonstrate low albumin, high creatinine and low CRP.

In contrast, participants in Cluster 11, who had the worst survival outcomes, (95 admission episodes from 75 unique participants), were associated with poor baseline function, high baseline cognitive function but biochemically low albumin, urea, creatinine and CRP. During acute illness, they demonstrated relative minimal arousal deficits, low delirium severity, low scores of illness severity and were generally younger. Cluster 2, made up of 40 admission episodes from 36 unique participants, also demonstrated bad survival outcomes. This cluster was associated with poor baseline cognition and baseline physical function, but high delirium severity and attentional deficit scores when unwell, low albumin without particularly raised inflammatory markers.

7.3.5 Length of stay prediction

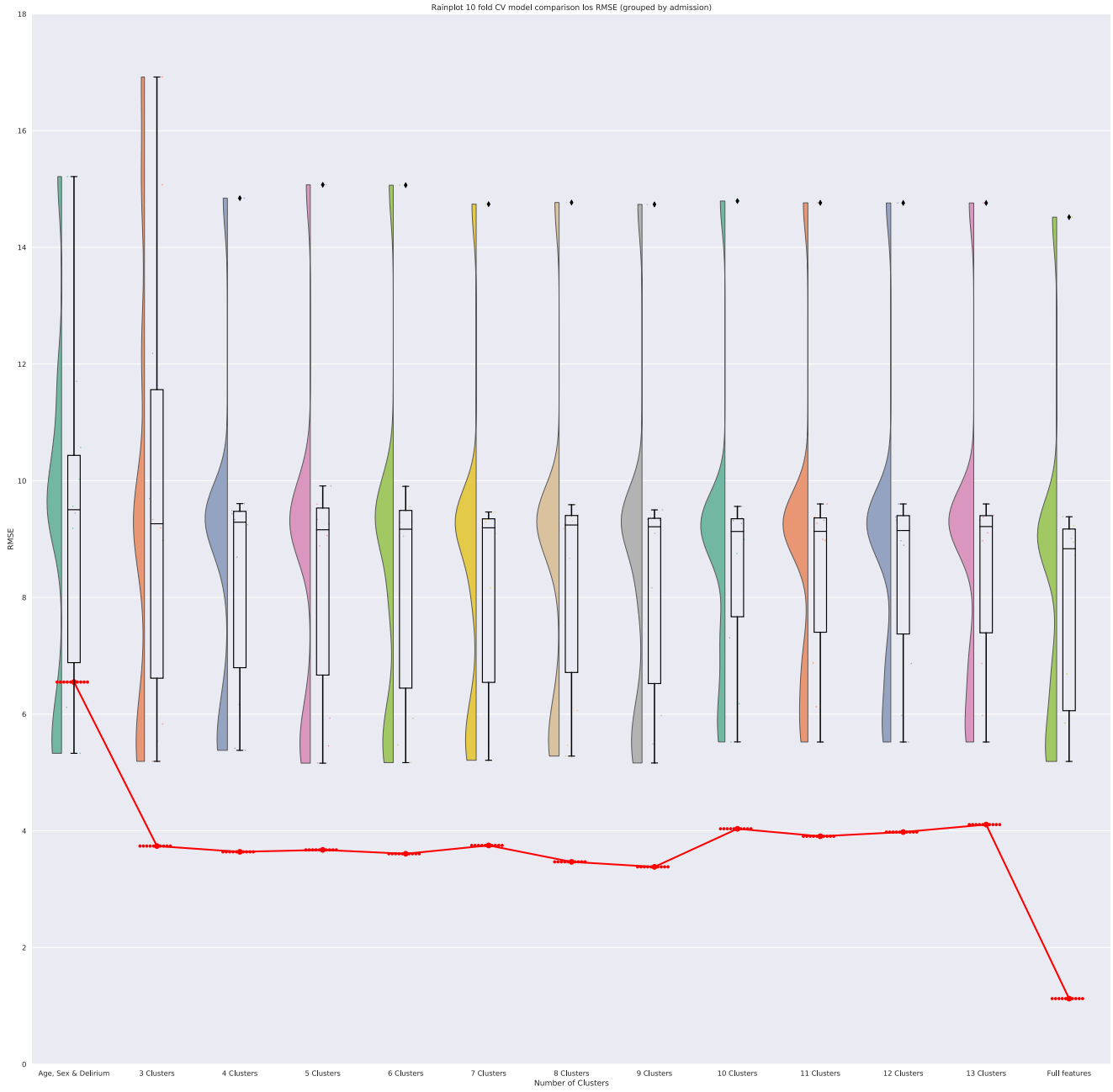
A general trend of training and testing performance demonstrated that the baseline model using age, sex and delirium was the least predictive model for length of stay (test RMSE = 6.4) (fig 7.7). When delirium was replaced by cluster designation, test predictive performance improved but with minimal differences between using 3 to 13 clusters. Test RMSE ranged between 3.8 and 4.1. The best performing mean train and test predictions were made using full features (test RMSE = 1.1). The train-cross-validation models performed consistently more poorly than the test partitions, likely a sequalae of random train-test split and reflective of the small number of observations, particularly in the test partition.

The two patient groups with good survival outcomes, clusters 1 and 3, correspondingly also had the two shortest lengths of stay (table 7.1). However, the two groups with poor survival outcomes had relatively divergent admission lengths: cluster 11, patients with poor baseline function but good baseline cognition, had relatively short lengths of stay. In contrast, in Cluster 2, patients in whom had both poor baseline cognition and function, had admission episodes over three times longer on average.

Table 7.1: Length of stay by cluster

Cluster	Number of participants	Length of stay (mean, IQR) (days)
1	55	1.9 (1-2)
2	36	9.7 (6 - 11)
3	33	2.7 (1-3)
4	3	7.4 (4-10)
5	30	8 (3-11)
6	9	5.6 (3-7)
7	15	3.3 (1-4)
8	3	48.7 (39-55)
9	1	6
10	13	20.6 (11-26)
11	75	3 (1-4)

Fig 7.7: Rainplot comparing length of stay predictions (RMSE) by baseline (age, sex, delirium status), with additional clustering and full features



Rainplots and confidence intervals generated from ten-fold cross validation folds.

Red line denotes C-index on held out test partition

7.4 Conclusions

7.4.1 Presence of structure during acute decompensation of older patients

This pilot study, performed with a relatively limited number of observations, shows discernible structure in the patterns of decompensation during acute illness in older people. The clinical and biochemical distributions of acute changes are not random but likely clustered into subtypes of prognostic significance, as demonstrated with survival and length of stay. Optimal clusters definition requires knowledge of, but not limited to, a patient's premorbid baseline cognition and functional deficits, as well as cross-sectional clinical features including cognitive performance, neuropsychiatric features and biochemical measures.

Despite the preliminary nature of these analyses, it is already evident that there are finer granularities of detail to be captured at baseline and during acute illness. Identifying poor baseline function, low albumin, urea and creatinine in clusters with the poorest survival profiles, essentially highlights the importance of detailed multimodal definition of frailty in modelling mortality risk. The biochemical profile of a low muscle mass state correlated with poorly mobile functional state, is in keeping with the predictive importance of sarcopenia from ML models in Chapter 6.

Conventional measures of raised inflammation do not necessarily correlate to a poor survival outcome during acute illness. It is possible the frailest patients may not physiologically generate the same proinflammatory state in response to an infectious stimulus as individuals with more robust premorbid baselines. However, it should also be considered whether a significant proinflammatory insult is necessary for death of the frailest patients – in this cohort, it is possible that mortality risk became overwhelmingly driven by the poor baseline, rendering conventional markers of illness insignificant, whether or not they are raised. The patient may have decompensated without generation of any significant systemic inflammation, reflecting minimal resilience to physiological insults in frailest patients. In turn, this implies acute illness is important for survival prediction only if a patient's baseline function and cognition exceed a minimum threshold.

For length of stay, poorer long-term mortality outcome does not necessarily equate to longer admission. The disparity between the two clusters with poor survival outcomes but significantly different inpatient durations suggest baseline cognition is the most significant risk to admission length, with baseline physical function of minimal contribution. This raw analysis should be interpreted with respect to likely confounders – higher baseline cognition is likely to be associated with higher educational attainment, patient and family socio-economic status, with probable

already established social arrangements for care and greater personal ability to augment care if required. As a result, baseline cognition is unlikely to be the primary physiological mechanism towards a shorter length of stay. However, it may nonetheless be the most proximal indicator of contributing factors articulated above, hence offering a simple predictor with high clinical utility for personal, family and healthcare organisational planning.

7.4.2 Importance of clustering

This clustering approach was not designed to outperform to high-dimensional models for prediction. However, clustering can identify possible underlying mechanisms common across patients with specific outcomes. Empirically defining clusters could lead to our developing subtype-specific clinical management. At the same time, improved stratification of potentially common mechanisms and outcomes may refine patient selection for future therapeutic trials.

These preliminary findings emphasise the importance of interpreting acute illness in the context of baseline features, with more accurate prediction possible with increasing number and complexity of multimodal input. At the same time, they highlight the deficiencies of currently used terms such as the DSM-IV definition of delirium: a cross-sectional entity without objective comparison to baseline, biochemical or imaging inputs with only a narrow inclusion of cognitive factors without other clinical or psychiatric features. While the current definition of delirium may be sufficient to predict cognitive decline (Tsui, Searle et al. 2022), adding features may identify new subtypes. Broader and more detailed descriptions of acute illness, incorporating understanding of a patient's premorbid baseline, may be more useful than current terms used in acute illness than current DSM-IV definition of delirium, which reflects on cross-sectional acute confusional state alone. These novel definitions may have offer prognostic prediction, be more informative on therapeutic strategies, while better explaining underlying pathophysiological mechanisms. With greater clinical use, these new conceptual terms may supersede current labels such as delirium.

7.4.3 Strengths and limitations

The scale of cross-sectional phenomenology and biochemical data, linked to outcomes and baseline, is the largest dataset of acute illness phenomenology collected outside critical care. The data are high-dimensional and multi-modal, including prospectively collected detailed ascertainment of cognitive and physical functional baseline. This is essential for understanding the context of subsequent acute illness. The cross-sectional clinical features across the spectrum of neuropsychiatric dysfunction, motor phenomenology, contemporaneous physiological and biochemical measures of illness severity can all be linked to prospective follow-up of outcomes. In

addition, I employed techniques agnostic to current definitions and pre-defined paradigms on acute illness in older people or delirium, allowing for a data-driven novel approach to subtype definition.

Hierarchical clustering utilises no a priori inputs of the number of clusters, allowing the most clinically useful definitions to be articulated without prior prejudice from precedence. Prediction models were constructed using boosted trees machine learning techniques amenable to non-linear associations, autocorrelation of variables while applying weighting of feature importance as appropriate in the data structure.

My findings should be interpreted in the context of their limitations. Despite DELPHIC being the most extensive population dataset available, the number of inpatient days observations is still small relative to the number of derived clinical features after summary within an admission episode, particularly when multiple admissions are excluded for a participant in the testing split. There are features of the comparative survival rainplots suggestive of the limited size of the dataset. This includes the relatively wide variance between ten folds of cross-validation within the training set for the two most basic models, with narrower variation in training performance for the model using full features. Between prediction models, the test scores consistently outperformed mean performances in training partition for all models, again likely reflecting over-estimation of true performance due to the limited observations, particularly in the test partition. While the test C index trend across models allows comparisons between hierarchical models of increasing dimensionalities, the precise value of test performances may not be sufficiently precise to be definitive. I await external validation of our findings in an independent but harmonised cohort.

Second, the dataset lacks ascertainment of illness within the community, either before admission, after discharge or for participants who do not become admitted to hospitals. As a result, too few study participants have clinical features data for predictive models of cognitive decline. Lastly, while the study utilises validated measures of delirium, these scales are limited in their outputs being multi-value ordinal data, with only blood data truly continuous in nature. During clustering analyses, a decision was thus required to use Euclidean approaches while recognising the limitations in applying to ordinal data. However, despite suboptimal operationalisation of ordinal data, the demonstrable clinical utility somewhat validates this pragmatic approach.

Third, this chapter provides a proof of concept that patterns of decompensation in older people during acute illness can be informatively identified using longitudinal phenomenology and biochemical abnormalities. In a relatively small sample size, approaches that seek to test hypotheses would be reasonably subject to concerns of underpowering. However, by instead

utilising supervised machine learning in cross-validated train and held-out test partitions to measure clinical predictive fidelity, evaluative emphasis is shifted to metrics of test model performance, instead of minimum power for regression analysis. Such an alternative approach is inevitably accompanied by alternative concerns, namely hyperparameter optimisation and overfitting, which nonetheless require ever larger bigger datasets to improve performance as well as external data to demonstrate generalisability.

8 Conclusions

Chapter Outline

- Strengths of the thesis
- Main thesis findings
- Ongoing limitations and challenges in research of acute illness in older people
- Questions for future research
- Novel approaches to future studies
- The future: a hypothesised case study

8.1 What are the strengths of this thesis?

As identified from chapter 3, there were multiple remaining evidence gaps within our knowledge of acute illness decompensation in older patients and in particular, delirium: first, although delirium period prevalence and symptomatic duration had been reported in elective surgical and intensive care settings, the epidemiology of delirium in medical older patients outside a critical care environment was poorly understood. The prevalence of delirium in older patients presenting to medical wards on the day of admission was unknown, as was the prevalence of unresolved delirium being discharged into the community.

Second, the relative contributions to adverse sequelae after delirium were unclear. It was unknown whether delirium is simply an unmasking of underlying cognitive impairments or whether delirium is toxic per se, with possible dose dependent adverse consequences. There were inconsistencies within literature on the clinical phenotype that the term “delirium” referred to, as a result of shifting reference standards, simplistic cross-sectional diagnosis of delirium as a binary yes/no (or hyper/hypo/mixed) constructs, difficulties in operationising screening tools and insufficient emphasis on arousal deficits within diagnostic instruments. Fourth, the performances of long-term outcomes prediction models in older people were inconsistent, generalised poorly and strongly dependent on health-care settings. Last, there had been limited objective articulation of patterns of illness decompensation among older patients, despite clinical experience advocating the presence of phenotypic subtypes, with a status quo relying on over-simplistic descriptions such as hyper or hypoactive delirium.

Here I will introduce how findings from this thesis have advanced previous research. For each finding, I will describe limits of previous evidence, summarise new findings from this thesis,

followed by how I have advanced understanding within the research field as well as their clinical implications for daily practice.

8.1.1.1 Population-based associations of delirium and long-term cognition

Incident delirium was known to be associated with increased mortality risk and greater rate of cognitive decline, independently of Alzheimer's pathology (Davis, Muniz Terrera et al. 2012). However, robust adjustment for prospectively ascertained baseline cognitive status and delirium ascertainment were lacking, frequently reliant on retrospective collection alone. Few studies were performed outside critical care settings and even fewer involved community-based sampling (Devore, Fong et al. 2017). Studies presumed linear associations between delirium and cognitive outcomes, with more complex analyses limited by lack of statistical power.

My findings first described the epidemiology of incident delirium in a predominantly medical population, as well as the prevalence of delirium on the day of admission, as well as significant number of patients being discharged with residual delirium into the community. In addition, I showed that poorer baseline cognition was associated with longer delirium duration, but those with the highest and lowest baseline cognition experienced the most severe symptoms in the event of delirium. Delirium was associated with greatest cognitive decline in those with highest baseline cognition. However, while better baseline cognition was protective against future cognitive impairment and death, the advantage of better baseline cognition are lost when exposed to high delirium burden.

These findings highlight the importance of considering non-linear associations when modelling acute illness in the oldest-old, including interactions of acute illness with the baseline state. Clinically, this emphasises the need for accurate baseline cognition assessment, such as the IQCODE (Blandfort, Gregersen et al. 2020), to identify the population cohort with the "most to lose", namely those with best baseline cognition. The high burden of cognitive impairment and mortality after incident delirium has implications for health services, for example, establishing delirium follow-up services for patients at highest risk of adverse outcomes. These epidemiological findings may encourage the inclusion of baseline cognition and delirium burden to illness severity scales such as early warning scores.

8.1.1.2 High-dimensional multimodal machine learning is feasible and accurate in oldest-old patients: a demonstration with long-term mortality prediction

Previously, predictive accuracy of mortality risk was poor for oldest-old patients, mainly only estimating short-term prognosis with limited generalisability across healthcare settings and

cultures (Curtin, Dahly et al. 2019). Models utilised limited numbers of within-patient features such as sex and comorbidities, one-year death risk from population-based life tables, and hospitalisation-specific factors such as number of admissions in one year, admission urgency, clinical service referred to and laboratory-based acuity scores (Curtin, Dahly et al. 2019). Their utility was thus limited beyond the specific clinical settings they were validated in, with wide performance variations across population and age groups, while some models were dependent on background information not readily available in acute settings (van Walraven and Forster 2017).

In this thesis, I showed how highly accurate, long-term mortality prediction is possible by using high dimensional, multimodal approaches while using cross-sectional routine data without *a priori* knowledge, producing prediction models agnostic to healthcare cultures and settings. This approach demonstrated proof-of-concept that constructing accurate prediction models in patients with interacting multi-morbidities is feasible. The implication for research includes greater emphasis on machine learning for modelling complex multifactorial biology, such as oldest patients, and the need to account for potentially non-linear associations in prediction modelling. Clinically, these findings offer significantly improved individualised prognostication, allowing better patient and family communications while guiding clinicians on the most appropriate management course. Identifying patients at the highest risk for health services allows allocation of finite healthcare resources to prioritise those at greatest risk of deterioration. Objective identification of sarcopenia and reduced bone density, as well as intracranial pathologies, being significant contributors to long-term mortality risk, again highlight the need for comprehensive geriatric assessments when managing older patients during acute illness.

8.1.1.3 Inference of intracranial brain and extracranial skull structures

Previously, contributors to mortality risk in older patients had been restricted to basic patient characteristics, such as age, sex, cancer status, and healthcare setting-specific factors such as the acuity of medical needs, rather than anatomical or biochemical factors. Features were frequently binary or categorical, limiting their power for predictive and inferential models.

My analyses demonstrate the relevance of including structural neuroimaging to improve prediction models. Using CT scans segmented with adapted MRI algorithms makes it possible to achieve a high degree of information from a relatively accessible imaging modality. The overall approach appears to identify targets for associated pathophysiology. Neuroanatomical regions such as the anterior cingulate and angular gyri as specific areas of interest, indicating motor and autonomic functions may contribute to mortality risk. They also implicate mechanisms associated with musculoskeletal abnormalities commonly referred to in sarcopenia, osteoporosis and presenting

non-specifically as “falls”. Adding structural intracranial and extracranial neuroimaging would augment current definitions of frailty, improving the clinical utility of the term “frailty” with its additional prognostic information, highlighting potential targets for further mechanistic research. However, causal directions in terms of whether central neural dysfunction directly results in motor deficits, or conversely if long-term immobility results in neural adaption of which central motor degeneration is a consequence, remains unclear.

8.1.1.4 Patterns in acute illness of older people

Although cognitive and functional decompensation during acute illness were already well recognised in older patients, the best-utilised quantification related to subjective measures of mobility deficits. In daily clinical practice, delirium was frequently conceptualised as a binary diagnosis, with colloquial subtypes primarily based on motor phenomenology.

Despite the limited numbers of initial observations available for analyses, preliminary findings within this thesis demonstrated underlying data structures during acute illness of older patients within cross-sectional clinical, biochemical, and premorbid baseline features. I showed distinct clinical patterns within clusters, with differing adverse event outcomes for mortality risk and lengths of stay.

The next step requires validation of these preliminary results on an external, independent dataset to demonstrate generalisability. At the same time, as shown in chapter 6, the richness of clinical data can be augmented by including additional modalities as they become available, such as neuroimaging, potentially improving cluster definitions.

Pragmatically, my findings may offer a novel term to describe acute illness in older people, contextualising the acute presentation by encompassing pre-morbid cognitive and baseline functional performances. This may be used in conjunction with, or perhaps even supersede if more clinically useful, the current cross-sectional definition of delirium. The benefits of subtyping range from significantly improved stratification when planning clinical therapeutic trials to highlighting of common pathophysiological pathways, which may be targets for potential novel research and pharmacological or non-pharmacological therapies.

8.1.2 Datasets

The size and granularity of both datasets were particular strengths of this thesis: DELPHIC is, to my knowledge, the largest dataset currently collated with cognitive ascertainment in older patients with baseline, long-term follow-ups and incident acute illness clinical and biochemical features.

The UCLH admissions cohort is one of the largest available multimodal datasets, comprising of cognition, blood, neuroimaging, and mortality outcomes. The consecutive nature of clinical data capture minimised recruitment bias inherent to research cohorts while implementation of novel CT segmentation algorithms allowed accurate delineation intracranial and extracranial of contributions towards predictive model fidelity.

In addition, the study design for DELPHIC was particularly unique: contextualisation of acute illness against an accurate pre-morbid baseline allowed differentiation between contributions of underlying cognitive and functional impairments, and toxic consequences of delirium per se. Standardised baseline and outcome cognitive measures, with discrete fluency and memory ascertainment, allowed direct long-term longitudinal comparisons with domain-specific detail. Inclusion of multiple baseline factors particularly important to older people, such as care input, mobility aid requirements, commonly missing from cohort studies, offered a comprehensive assessment of baseline frailty to optimise covariate adjustments. Last, delirium ascertainment during acute illness episodes was particularly rich and unique: specific inclusion of arousal features, assessment using a severity scale, consecutive daily assessments to capture phenotypic fluctuations, concurrent inclusion of functional activities and mobility impairments, all provided a multifaceted articulation of acute illness and delirium.

8.2 Ongoing limitations and challenges across the research field

There are significant challenges to the study of decompensation in older patients during acute illness. Some issues may be mitigated by thorough and innovative study planning. However, others are challenges either specific to or particularly prevalent in this patient cohort, which must be considered when interpreting any future study results. I will describe research challenges related to sampling and study design, outcome ascertainment and choice of dependent variables. For each, I will first highlight how the DELPHIC and UCLH acute illness studies have been novel in mitigating some of these limitations, advancing previous knowledge through more robust methodology. However, I will also articulate how other limitations remain yet unaccounted for within my thesis, and how they can be minimised in the future.

8.2.1 Study design and sampling

Selection bias is a common issue for both epidemiological and machine learning studies, involving patients of all ages, potentially resulting in models that generalise poorly to previously unseen data in new healthcare cultures and settings. Consequently, in epidemiological studies, associations and effect sizes demonstrated may only be applicable in a relatively narrow population from which

the sample was drawn, while in machine learning models, performances become less accurate when applied to different population cohorts. External validation is the most robust method to confirm validity and robustness of an epidemiological or machine-learning model. However, availability of sufficiently similar external datasets, in independent geographic and healthcare setting with harmonisable outcomes and features for comparison, are particularly scarce in older people research.

The study design of DELPHIC aimed to limit selection bias by utilising a population-based sampling frame. Potential participants were excluded only for practical reasons such as ability to consent due to language, capacity, or hearing. There were active efforts to recruit patients from residential institutions and over-sampling patients known to the memory service to include previously neglected population groups. Similarly, the UCLH acute illness dataset included consecutively admitted patients to the medicine for the elderly service, regardless of reason for admission.

Nonetheless, limitations remain for both studies. In DELPHIC, patients for whom English is not their first language and those cognitively impaired without a named next of kin were inevitably under-represented. In addition, findings from both studies may generalise less well to rural populations: participants in DELPHIC live in an urban environment with higher educational attainment and more affluent socioeconomic status compared to the national average. Similarly, the UCLH dataset consisted of patients acutely admitted from a central London population that overlaps with the DELPHIC sampling frame (Goodyer, Mah et al. 2022). As a result, rural populations and those with lower levels of formal educational attainment are not as well represented in the studies within this thesis.

Third, there is an accepted logistical difficulty in designing studies across primary, secondary and community care. Both studies in this thesis included inpatients only, with no ascertainment for decompensation that may present as delirium and treated in primary care alone. However, delirium is most often present at the point of hospital admission (e.g. it developed beforehand), and in DELPHIC, 40% of patients with delirium were still delirious on discharge. Accurate ascertainment of acute illness is therefore limited to secondary care capture alone. As a result, study findings may potentially generalise poorly for patients in the community. In addition, the true duration, severity and hence cumulative exposure of an illness and its associated clinical syndromes such as delirium, cannot accurately measured, resulting in less accurate dose-dependent models of acute illness, delirium and their consequent adverse events.

8.2.2 Outcome ascertainment

Heterogeneity in outcomes is common for older people. This makes it harder to demonstrate true underlying associations with dependent variables and reduces accuracy of prediction models. The cardinal outcomes for acute decompensation generally include mortality, long-term cognitive impairment, delirium incidence, length of stay and increased level of care. Mortality is an easily defined and accurate binary outcome. This is not the case for length of stay or increased level of care, which in addition to organic factors, are determined by a range of non-organic factors such as premorbid social organisation, capacity of local social services, availability of district nursing, delivery of functional equipment and rehabilitation capacity.

For delirium, target definition remains challenging. Firstly, even with the gold standard DSM-IV criteria, difficulty in operationalising has resulted in a range of possibilites, each with its own subjective interpretation. Regularly used tools in research include CAM-ICU, DRS and MDAS (Inouye, van Dyck et al. 1990, Rockwood, Goodman et al. 1996, Breitbart, Rosenfeld et al. 1997), with poor ability to be directly harmonised and hence robustly comparable. In addition, most accepted tools may be insufficient to capture reference-standard delirium. Even in the case of the DELPHIC study, with dedicated research staff providing daily inpatient assessments and was the most detailed and intensive delirium ascertainment of any cohort to date, records of features fluctuations were limited to retrospective corroborative accounts as formal assessments were performed only once in 24 hours.

There remains no consensus method for quantifying delirium dose. This thesis proposed the relatively crude product of delirium duration and daily severity. However, it is already recognised that the current measures of delirium duration are suboptimal, particularly when studies are restricted to secondary care alone. First, the lack of ascertainment outside of inpatient care would exclude uncaptured pre and post-hospital burden, or any delirium at all in patients not admitted to hospital. Second, if only inpatient care is included within studies and considering significant number of patients are discharged still with delirium by DSM-IV criteria, delirium duration becomes increasingly entwined with length of stay and its associated range of non-organic factors. Measures of the delirium severity component to burden is similarly challenging, even if using a single metric such as MDAS. Despite being a scale for delirium, a number of items will likely score in patients with dementia alone but without delirium. As a result, more innovative operationalisation is required, such as additional thresholding to only include DSM-IV positive MDAS scores or subcortical items without those that overlap with a dementia only diagnosis, to refine existing tools.

Next, delirium will require articulation in greater granularity than an incidentally binary diagnosis. The findings in Chapter 7 raise questions about different clinical, cognitive and biochemical patterns, at baseline and at the point of decompensation, resulting in differential adverse outcome risks. Cross-sectionally, describing delirium alone in greater granularity may no longer be the most clinically useful approach. Instead, aiming for a broader term that includes a greater range of features, incorporating those currently used to describe delirium, may have greater utility.

Longitudinally, it is currently assumed that presence of delirium is an equal construct for all patients, at all temporal points of their delirium course. However, it remains unclear how delirium results in toxic effects on long-term adverse outcomes. Given the fluctuating nature of the syndrome and potential reversibility of symptoms, it seems unlikely that delirium pathophysiology is entirely underpinned by neuronal loss. Therefore, a reversible intermediate delirium state, perhaps difficult to delineate using clinical examination alone from an irreversible unidirectional deterioration trajectory, awaits articulation. This deep phenotyping will likely encompass a combination of baseline and incidental clinical features, corroborated biomarkers, neural network ascertainment, and multimodal structural and functional neuroimaging. This approach will improve rigour of definitions and add greater consistency to entities such as persistent delirium, syndromal delirium, delirium superimposed on dementia and repeated punctuated episodes of delirium.

8.2.3 Lack of multimodality, granularity and prosepective capture in covariate ascertainment

Challenges in ascertainment of dependent variables limit optimal epidemiological model adjustment, machine learning model predictive power and ability to describe significant interactions. Dependent variables are commonly retrospective, observer-dependent and subjective, resulting in significant interrater variability and inherently inaccurate or incomplete capture. In addition, the binary or ordinal nature of data, such as presence of symptoms and MDAS items respectively, frequently limits power within modelling. Even when continuous measures have been available, for example, physiological measures such as heart rate and blood pressure, there has been a pragmatic trend towards thresholded measures at sparse time intervals for ease of manual calculation (e.g the National Early Warning Score). Despite the rapid technological advances in availability of wearables, motion and geographic sensors, these have been slow to be incorporated into routine use for data collection in large cohorts, with currently unrealised potential of large volumes of longitudinal, highly granular time-series data.

Specifically in the study of decompensation during acute illness, accurate articulation of illness severity is problematic in older people. It is well recognised that routinely used inpatient physiological parameters are suboptimal for articulating illness severity in older patients: clinical

experience informs us that temperature spikes do not always occur even in severe infections. Markers of acute inflammation such as white cell count and C-reactive protein (CRP) may not rise to the extent as would be expected in younger patients. Appropriate adjustment for the amplitude of the insult is essential for modelling decompensation during acute illness. This is unlikely to be practical nor useful to be a measurement of the dose of the insult itself. As witnessed during the Covid pandemic, peak of symptoms and decline occurred approximately 7 days after the infection, differential outcomes in older people are unlikely to represent the size of the viral exposure received, but rather more downstream mechanisms that reflected varying degrees of inflammatory responses.

Lastly, as briefly explained as a corroborative tool for target ascertainment, the use of multi-modal data sources remains lacking. Our findings in chapter 6 highlight the importance of multimodality in maximisation of accuracy in predictive algorithms and no doubt, extends for adjustment in epidemiological models. Practical and funding challenges and a lack of validated consensus operational protocols are well-recognised in large cohorts. Confused patients are frequently unable to stay sufficiently still for artefact-free imaging. While functional imaging would likely provide novel insights into hypothesised dysfunctional brain networks during acute illness, findings may be confounded by sedating drug effects necessary for sufficient image capture. Recognised resting and task-based imaging protocols would be limited by poor engagement during acute confusional states. Meanwhile, the cost and availability to process and analyse highly specialised tests such as genomics, proteomics or metabolomics await to be overcome.

8.2.4 Questions for future research

This final sub-section identifies the next urgent research questions extending from the novel findings from this thesis. Because of the previously described heterogeneous nature of outcomes such as length of stay and increased care needs, I will focus these questions on delirium and mortality.

8.2.4.1 Epidemiology of delirium

An immediate question is to further characterise the broad heterogeneity of delirium by longitudinal cognitive and clinical features and biochemical profiles, as explored in chapter 7 and awaiting external validation. Further work is required to address how the temporal course of delirium impacts potentially adverse outcomes.

1. Can delirium or acute illness be further classified into further adverse outcome-driven subtypes?

2. Is persistent delirium as toxic a driver towards cognitive decline and mortality as incident resolved delirium? At what duration does delirium become persistent based on adverse outcomes? Does persistent delirium demonstrate the same reversibility profile for patients with different baseline cognition?

3. Are punctuated repeated delirium episodes independent of their adverse effects, or do they produce adverse outcomes more severe than their individual episodic components?

4. Is post-delirium cognitive decline immediately irreversible in all patients, or does it progress via an intermediate reversible state?

5. If there is a period of intermediate reversibility, does the length of a window of reversibility differ for patients of varying baseline cognition?

8.2.4.2 Pathophysiological substrate

Understanding the neural substrate of post-delirium cognitive impairment is an urgent frontier in delirium research. The differential effects of incident delirium on varying cognitive baselines suggest different underlying neuropathology depending on premorbid cognitive state. Within the acute episode, the fluctuating nature of delirium, with periods of deterioration and transient improvement, seem poorly explained by neuronal death alone, which would inherently result in unidirectional deterioration. Defining the culprit mechanism will add an entirely new dimension to the understanding of delirium, in terms of better defining delirium subtypes, prognoses and appropriateness of treatment streams. A more likely hypothesised candidate mechanism would be disruption of functional brain networks: evidence of global network oscillation dysfunction from EEG studies from patients with delirium have demonstrated lowing or dropout of the posterior dominant rhythm, generalized theta or delta slow-wave activity, poor organisation of the background rhythm (Jacobson and Jerrier 2000).

1. Is there evidence of progressive functional network dysfunction in the initial presentation of delirium, in the short-term post delirium presentation and in the long-term when cognitive impartment is clinically demonstrable?

    a. Are there changes demonstrable in global measures such as EEG features, extraneuronal field potentials such as from deep brain oscillatory recordings?

    b. Can they be demonstrated using resting state and evoked tasks on functional neuroimaging?

c. Can drug effects be sufficiently separated from organic pathological signals?

d. Do features of network dysfunction differ depending on baseline cognition?

e. Do biomarkers of neuronal damage, such as neurofilament light, correlate to network abnormalities and are these patterns different for varying cognitive baselines?

f. Is network dysfunction caused by local or global pathogenic activity or loss of global coherence between essential brain regions?

g. Can dynamic modelling demonstrate causality between focal anatomical pathologies and clinical phenotype?

2. How does any functional dysfunction mechanistically result in neuronal death?

    a. Is there an analogous comparison with epileptiform activity, in which chronic pathological intracerebral discharge would result in neuronal damage?

    b. Does chronic disconnection of brain regions and subsequent hypoactivity progress towards neuronal loss?

8.2.4.3 Individualised prediction

Building on the anticipated advances in better delirium definition and patient stratification by mechanisms and outcomes, there should be aims to extrapolate findings from populations to individual patients.

1. What is an individual's risk of delirium incidence for a defined anticipated insult, such as elective surgery? What is the anticipated duration, severity and burden?

2. Following on from an episode of delirium, what is the individual patient's risk of mortality and cognitive decline of pre-determined amplitude?

3. What is the degree of anticipated reversibility of cognitive dysfunction for a patient with current delirium?

8.2.4.4 Therapeutics options

Clinical trials of potential therapeutic strategies can be planned in patients predicted to have or demonstrated to show potential reversibility of cognitive decline. Mechanism-agnostic therapies such as cognitive rehabilitation in the short-term after delirium and "brain training" maintenance

exercise would likely be strategies. However, confirmation of network dysfunction may raise further interest in potential specific therapies ranging from repurposing existing drugs, such as anti-epileptics to more interventions modulating brain networks, such as transcranial magnetic stimulation.

1. Can a randomised controlled trial of cognitive rehabilitation demonstrate therapeutic benefit in a pre-stratified population of patients with delirium showing features of likely cognitive reversibility?

2. Building on robust predictive models of cognitive decline, can stochastic therapeutic inference demonstrate signals of candidate drugs from existing formularies that may reduce long-term impairment for further larger trials?

3. Are focal or global network dysfunction demonstrable in patients with reversible delirium deficits modifiable by targeted experimental therapies such as transcranial magnetic stimulation?

8.2.5 Novel approaches to future studies

Answering these outstanding questions will no doubt require concurrent approaches using traditional epidemiology and machine learning methods. Relevant to all population cohort and machine learning studies, the most limiting current issue is not incorporating robust external validation. This needs harmonisable outcomes and predictive features across international collaborations to overcome local overfitting. Mutually validating models could also aim to generalise across languages and cultures. It will be crucial for future studies to be designed with closely collaborative national and international centres in simultaneous conjunction.

It is essential to anticipate the exponential need for computational power to study acute decompensation sufficiently. This would have to accommodate greater volumes of data and more computationally intensive techniques. Future studies will utilise novel real-time monitoring modalities to collect continuous scales and time series data in dense temporal epochs, replacing ordinal or binary single data such as from informant reporting. For example, this evolution could include: deployment of extracranial modalities to study cerebral networks (e.g. EEG and MEG); real-time data collection of specific measures such as wearable heart rates monitors, eye movements, and behavioural measures through infra-red motion detectors. Increasing digitalisation of healthcare and everyday life will generate unprecedented volumes of data, each able to provide clinically useful data if able to extract significant signals from inevitable noise.

Furthermore, incorporating devices outside typical medical modalities will offer an increasingly complete articulation of a patient's clinical phenotype, such as smart meters for electricity and water consumption, smart devices on appliances including refrigerators, kettles to describe stereotyped and abnormal behaviour. Unsupervised machine learning techniques such as deep neural networks will likely supersede supervised tree-based algorithms as datasets become larger and more complex, requiring projection across multiple manifolds.

The outcome variable will become increasingly detailed and articulated for epidemiological and machine-learning approaches. Biomarkers of irreversible neuronal damage and dementia pathology such as neurofilament light, amyloid ratios and phosphorylated tau could be more routinely integrated as outcome variables. It will be essential to utilise reversible markers of delirium mechanisms of any intermediate delirium state, either in isolation such as depressed oscillatory network frequency activity or resting state MRI abnormalities. The DELPHIC study establishes the absolute requirement of baseline ascertainment to measure cognitive decline robustly. At the same time, the definition of delirium itself will become increasingly harmonisable and comparable across studies, with greater agreement over definitions of duration, severity and composition of motor and non-motor clinical features.

Machine learning prediction models extend the findings of associations identified on a population-level epidemiology and allow individualised application at the patient level.

Incorporating an increasing number of modalities and feature dimensions, such as structural and functional neuroimaging, EEG, fNIRs and TMS, blood and CSF biomarkers of neuronal dysfunction and death, metabolomics, proteomics, transcriptomics and genomics can be used to construct increasingly complex and sophisticated unsupervised machine learning models. Using better-defined outcome ascertainment for delirium, care needs and mortality, such a model can inform patient and next of kin choices: what is the risk of delirium if a patient undergoes an elective procedure? If a patient in the community develops an illness of severity requiring hospitalisation, what is the risk of delirium when admitted to hospital? What is the likely duration, severity and overall burden if delirium occurs? What is the likely risk of significant cognitive impairment beyond X% of baseline and death following incident delirium? Quantifying these outcomes offers a novel dimension to informed patient choice when deciding on elective surgery, better advanced care planning for frail patients in the community, and appropriate early input from palliative services.

Prediction models for patients with incident delirium, in which the outcome is a probable intermediate reversible state, likely defined by markers of functional network dysfunction and lack of evidence of neuronal damage, would similarly transform acute care of patients with delirium.

These models would offer more objective methods of identifying patients most likely to benefit from currently available cognitive rehabilitation. Equally important, these models allow better stratification of patients most likely to benefit from future therapeutics for modifying the disease course of delirium and hence most appropriate to participate in clinical trials of candidate interventions.

Constructing accurate prediction models of potential recovery from delirium allows inferences to be made about the underlying neuroanatomical and functional inferences, as well as the potential for therapeutic inference. As pioneered in Chapter 6 of this thesis, neuroimaging features from a well-performing prediction model highlight potentially plausible contributory brain areas. Lastly, large datasets incorporating pharmacy data, in which reversibility of patients with delirium is available as an outcome, can be used to for therapeutic inference of candidate drugs for further investigations. For example, are patients with antiepileptics less likely to be predicted to experience irreversible cognitive impairment, and if so, what is the relative contribution of these drugs? The most likely candidates can be considered for feasibility pilot studies followed by potential larger clinical trials.

8.3 The Future

Mrs P is now an 88-year-old referred to the orthopaedic outpatient clinic by her GP, having experienced worsening bilateral knee pain for the past 18 months. Initial X rays demonstrated significant loss of joint space and osteophyte formation, consistent with severe osteoarthritic changes. Despite multiple trials of topical non-steroidal anti-inflammatory drugs and intra-articular steroid injections, the symptoms had progressed to the point her mobility was significantly restricted. She is now limited to short distances around her two-storey house with the aid of a walking stick and requires a stairlift. She attends with her son for a multidisciplinary assessment day to discuss the possibility of knee replacement surgery.

First, the surgeon, Miss G, following clinical examination and reviewing her images, agrees that surgery under a general anaesthetic would be appropriate. In addition to the intended benefits of the procedure, she describes the general risks of bleeding, infection and inability to complete the procedure successfully. She mentions that there is a risk of delirium for all patients undergoing a general anaesthetic and a surgical procedure. Mrs P and her son are alarmed at this prospect in particular – they recall an episode 15 years earlier when she was admitted to hospital with a chest infection, during which she had developed delirium. She recalls it being an extremely distressing experience as a patient, which her son also concurred to as well as a next of kin to witness. In addition, despite “recovering” from the confusion, both felt her memory had not been quite the

same. They were particularly surprised by how good her cognition had been prior to the chest infection. She also felt she had “slowed down” noticeably and now requires a carer twice a day to help her with dressing and washing.

Acknowledging this and encouraging both to discuss further, Miss G takes both the patient and her son to the adjacent consulting room, where a liaison geriatrician, Prof. D, awaits both. He first refers the patient for a new set of blood tests including an advanced multiomics and peripheral neurological biomarker panel, a structural and functional MRI of her head, an EEG and a battery of cognitive tests. Reviewing the completed investigations and utilising a now well-established algorithm, he counsels both on the likely risk of adverse events following a bilateral knee replacement under a general anaesthetic – Mrs P has a 23% risk of developing delirium at any point during this procedure and its recovery, 4% risk of mortality within 30 days and 36% risk at 2 years after this procedure, with an estimated 4 days length of stay as an inpatient. The risk of her requiring an increase to her current level of care is 3%. Should Mrs P develop delirium, however, there would be a 87% chance that reversibility to within 80% of her baseline cognition within 4 weeks – Prof D advises that the active treatments she would receive are cognitive rehabilitation in a dedicated post-delirium unit. In addition, she would be treated with a well-tolerated anti-epileptic medication to reduce the duration and severity of delirium, which has recently been identified from a large-scale inference study.

Mrs P emphasises again that it is the distress of delirium and “getting dementia after it” that she was initially most uneasy about but felt more reassured there are treatments to minimise the risk. At this point, Prof. D also introduces her to an ongoing research trial that she would be eligible for – the latest research suggests that manipulation of abnormal brain networks during delirium may be reversible or event prevented using techniques called fNIRS, median nerve stimulation and transcranial magnetiic stimulation – all of which are non-invasive.

Approaching the end of the assessment day, Mrs P and her son feel they are sufficiently informed of the potential risks of adverse effects after surgery and decided to proceed. They note that this is much different from when she first experienced illness and delirium 15 years ago, regarding how much doctors can tell her and how it can be treated. She is keen on enrolling in the research study and remarks that any intervention that can prevent or treat the delirium she had previously experienced would be life-changing for any patient.

References

"World Health Organization. World report on ageing and health. Geneva: World Health Organization; 2015.".

(ONS), O. o. N. S. (2018). "Living longer: how our population is changing and why it matters."

Ahmed, S., B. Leurent and E. L. Sampson (2014). "Risk factors for incident delirium among older people in acute hospital medical units: a systematic review and meta-analysis." Age Ageing **43**(3): 326-333.

Akunne, A., L. Murthy and J. Young (2012). "Cost-effectiveness of multi-component interventions to prevent delirium in older people admitted to medical wards." Age Ageing **41**(3): 285-291.

Ashburner, J. and K. J. Friston (2005). "Unified segmentation." Neuroimage **26**(3): 839-851.

Barber, R., I. McKeith, C. Ballard and J. O'Brien (2002). "Volumetric MRI study of the caudate nucleus in patients with dementia with Lewy bodies, Alzheimer's disease, and vascular dementia." J Neurol Neurosurg Psychiatry **72**(3): 406-407.

Blaiotta, C., P. Freund, M. J. Cardoso and J. Ashburner (2018). "Generative diffeomorphic modelling of large MRI data sets for probabilistic template construction." Neuroimage **166**: 117-134.

Blandfort, S., M. Gregersen, K. Rahbek, S. Juul and E. M. Damsgaard (2020). "The short IQCODE as a predictor for delirium in hospitalized geriatric patients." Aging Clin Exp Res **32**(10): 1969-1976.

Bone, A. E., N. Hepgul, S. Kon and M. Maddocks (2017). "Sarcopenia and frailty in chronic respiratory disease." Chron Respir Dis **14**(1): 85-99.

Breitbart, W., B. Rosenfeld, A. Roth, M. J. Smith, K. Cohen and S. Passik (1997). "The Memorial Delirium Assessment Scale." J Pain Symptom Manage **13**(3): 128-137.

Chen T, G. C. (2016). "XGBoost: A Scalable Tree Boosting System." KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining **August 2016**: 785-794.

Chester, J. G., M. Beth Harrington and J. L. Rudolph (2012). "Serial administration of a modified Richmond Agitation and Sedation Scale for delirium screening." J Hosp Med **7**(5): 450-453.

Clegg, A., M. Westby and J. B. Young (2011). "Under-reporting of delirium in the NHS." Age Ageing **40**(2): 283-286.

Collins, N., M. R. Blanchard, A. Tookman and E. L. Sampson (2010). "Detection of delirium in the acute hospital." Age Ageing **39**(1): 131-135.

Cook, S. E., M. Marsiske and K. J. McCoy (2009). "The use of the Modified Telephone Interview for Cognitive Status (TICS-M) in the detection of amnestic mild cognitive impairment." J Geriatr Psychiatry Neurol **22**(2): 103-109.

Crinion, J., R. Turner, A. Grogan, T. Hanakawa, U. Noppeney, J. T. Devlin, T. Aso, S. Urayama, H. Fukuyama, K. Stockton, K. Usui, D. W. Green and C. J. Price (2006). "Language control in the bilingual brain." Science **312**(5779): 1537-1540.

Cruz-Jentoft, A. J., G. Bahat, J. Bauer, Y. Boirie, O. Bruyère, T. Cederholm, C. Cooper, F. Landi, Y. Rolland, A. A. Sayer, S. M. Schneider, C. C. Sieber, E. Topinkova, M. Vandewoude, M. Visser and M. Zamboni (2019). "Sarcopenia: revised European consensus on definition and diagnosis." Age Ageing **48**(1): 16-31.

Curtin, D., D. L. Dahly, M. van Smeden, D. P. O'Donnell, D. Doyle, P. Gallagher and D. O'Mahony (2019). "Predicting 1-Year Mortality in Older Hospitalized Patients: External Validation of the HOMR Model." J Am Geriatr Soc **67**(7): 1478-1483.

Dani, M., L. H. Owen, T. A. Jackson, K. Rockwood, E. L. Sampson and D. Davis (2018). "Delirium, Frailty, and Mortality: Interactions in a Prospective Study of Hospitalized Older People." J Gerontol A Biol Sci Med Sci **73**(3): 415-418.

Davis, D., S. Richardson, J. Hornby, H. Bowden, K. Hoffmann, M. Weston-Clarke, F. Green, N. Chaturvedi, A. Hughes, D. Kuh, E. Sampson, R. Mizoguchi, K. L. Cheah, M. Romain, A. Sinha, R. Jenkin, C. Brayne and A. MacLullich (2018). "The delirium and population health informatics cohort study protocol: ascertaining the determinants and outcomes from delirium in a whole population." BMC Geriatr **18**(1): 45.

Davis, D. H., L. E. Barnes, B. C. Stephan, A. M. MacLullich, D. Meagher, J. Copeland, F. E. Matthews and C. Brayne (2014). "The descriptive epidemiology of delirium symptoms in a large population-based cohort

study: results from the Medical Research Council Cognitive Function and Ageing Study (MRC CFAS)." BMC Geriatr **14**: 87.
Davis, D. H., S. H. Kreisel, G. Muniz Terrera, A. J. Hall, A. Morandi, M. Boustani, K. J. Neufeld, H. B. Lee, A. M. Maclullich and C. Brayne (2013). "The epidemiology of delirium: challenges and opportunities for population studies." Am J Geriatr Psychiatry **21**(12): 1173-1189.
Davis, D. H., G. Muniz Terrera, H. Keage, T. Rahkonen, M. Oinas, F. E. Matthews, C. Cunningham, T. Polvikoski, R. Sulkava, A. M. MacLullich and C. Brayne (2012). "Delirium is a strong risk factor for dementia in the oldest-old: a population-based cohort study." Brain **135**(Pt 9): 2809-2816.
Davis, D. H., G. Muniz-Terrera, H. A. Keage, B. C. Stephan, J. Fleming, P. G. Ince, F. E. Matthews, C. Cunningham, E. W. Ely, A. M. MacLullich and C. Brayne (2017). "Association of Delirium With Cognitive Decline in Late Life: A Neuropathologic Study of 3 Population-Based Cohort Studies." JAMA Psychiatry **74**(3): 244-251.
De Buyser, S. L., M. Petrovic, Y. E. Taes, K. R. Toye, J. M. Kaufman, B. Lapauw and S. Goemaere (2016). "Validation of the FNIH sarcopenia criteria and SOF frailty index as predictors of long-term mortality in ambulatory older men." Age Ageing **45**(5): 602-608.
Detroyer, E., P. M. Clement, N. Baeten, M. Pennemans, M. Decruyenaere, J. Vandenberghe, J. Menten, E. Joosten and K. Milisen (2014). "Detection of delirium in palliative care unit patients: a prospective descriptive study of the Delirium Observation Screening Scale administered by bedside nurses." Palliat Med **28**(1): 79-86.
Devore, E. E., T. G. Fong, E. R. Marcantonio, E. M. Schmitt, T. G. Travison, R. N. Jones and S. K. Inouye (2017). "Prediction of Long-term Cognitive Decline Following Postoperative Delirium in Older Adults." J Gerontol A Biol Sci Med Sci **72**(12): 1697-1702.
Di Bari, M., D. Balzi, A. T. Roberts, A. Barchielli, S. Fumagalli, A. Ungar, S. Bandinelli, W. De Alfieri, L. Gabbani and N. Marchionni (2010). "Prognostic stratification of older persons based on simple administrative data: development and validation of the "Silver Code," to be used in emergency department triage." J Gerontol A Biol Sci Med Sci **65**(2): 159-164.
Drame, M., N. Jovenin, J. L. Novella, P. O. Lang, D. Somme, I. Laniece, T. Voisin, P. Blanc, P. Couturier, J. B. Gauvain, F. Blanchard and D. Jolly (2008). "Predicting early mortality among elderly patients hospitalised in medical wards via emergency department: the SAFES cohort study." J Nutr Health Aging **12**(8): 599-604.
Ely, E. W., R. Margolin, J. Francis, L. May, B. Truman, R. Dittus, T. Speroff, S. Gautam, G. R. Bernard and S. K. Inouye (2001). "Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU)." Crit Care Med **29**(7): 1370-1379.
European Delirium Association & American Delirium Society (2014). "The DSM-5 criteria, level of arousal and delirium diagnosis: inclusiveness is safer." BMC Med **12**: 141.
Fischer, S. M., W. S. Gozansky, A. Sauaia, S. J. Min, J. S. Kutner and A. Kramer (2006). "A practical tool to identify patients who may benefit from a palliative approach: the CARING criteria." J Pain Symptom Manage **31**(4): 285-292.
Fong, T. G., D. Davis, M. E. Growdon, A. Albuquerque and S. K. Inouye (2015). "The interface between delirium and dementia in elderly adults." Lancet Neurol **14**(8): 823-832.
Gibb, K., A. Seeley, T. Quinn, N. Siddiqi, S. Shenkin, K. Rockwood and D. Davis (2020). "The consistent burden in published estimates of delirium occurrence in medical inpatients over four decades: a systematic review and meta-analysis study." Age Ageing **49**(3): 352-360.
Girard, T. D., M. C. Exline, S. S. Carson, C. L. Hough, P. Rock, M. N. Gong, I. S. Douglas, A. Malhotra, R. L. Owens, D. J. Feinstein, B. Khan, M. A. Pisani, R. C. Hyzy, G. A. Schmidt, W. D. Schweickert, R. D. Hite, D. L. Bowton, A. L. Masica, J. L. Thompson, R. Chandrasekhar, B. T. Pun, C. Strength, L. M. Boehm, J. C. Jackson, P. P. Pandharipande, N. E. Brummel, C. G. Hughes, M. B. Patel, J. L. Stollings, G. R. Bernard, R. S. Dittus and E. W. Ely (2018). "Haloperidol and Ziprasidone for Treatment of Delirium in Critical Illness." N Engl J Med **379**(26): 2506-2516.
Goldberg, T. E., C. Chen, Y. Wang, E. Jung, A. Swanson, C. Ing, P. S. Garcia, R. A. Whittington and V. Moitra (2020). "Association of Delirium With Long-term Cognitive Decline: A Meta-analysis." JAMA Neurol **77**(11): 1373-1381.

Goodyer, E., J. C. Mah, A. Rangan, P. Chitalu, M. K. Andrew, S. D. Searle, D. Davis and A. Tsui (2022). "The relative impact of socioeconomic position and frailty varies by population setting." Aging Med (Milton) **5**(1): 10-16.
Gross, A. L., R. N. Jones, D. A. Habtemariam, T. G. Fong, D. Tommet, L. Quach, E. Schmitt, L. Yap and S. K. Inouye (2012). "Delirium and Long-term Cognitive Trajectory Among Persons With Dementia." Arch Intern Med **172**(17): 1324-1331.
Gunther, M. L., A. Morandi, E. Krauskopf, P. Pandharipande, T. D. Girard, J. C. Jackson, J. Thompson, A. K. Shintani, S. Geevarghese, R. R. Miller, 3rd, A. Canonico, K. Merkle, C. J. Cannistraci, B. P. Rogers, J. C. Gatenby, S. Heckers, J. C. Gore, R. O. Hopkins and E. W. Ely (2012). "The association between brain volumes, delirium duration, and cognitive outcomes in intensive care unit survivors: the VISIONS cohort magnetic resonance imaging study*." Crit Care Med **40**(7): 2022-2032.
Gusmao-Flores, D., J. I. Salluh, R. A. Chalhub and L. C. Quarantini (2012). "The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: a systematic review and meta-analysis of clinical studies." Crit Care **16**(4): R115.
Hayhurst, C. J., A. Marra, J. H. Han, M. B. Patel, N. E. Brummel, J. L. Thompson, J. C. Jackson, R. Chandrasekhar, E. W. Ely, P. P. Pandharipande and C. G. Hughes (2020). "Association of Hypoactive and Hyperactive Delirium With Cognitive Function After Critical Illness." Crit Care Med **48**(6): e480-e488.
Hendry, K., T. J. Quinn, J. Evans, V. Scortichini, H. Miller, J. Burns, A. Cunnington and D. J. Stott (2016). "Evaluation of delirium screening tools in geriatric medical inpatients: a diagnostic test accuracy study." Age Ageing **45**(6): 832-837.
Hsieh, S., S. Schubert, C. Hoon, E. Mioshi and J. R. Hodges (2013). "Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease." Dement Geriatr Cogn Disord **36**(3-4): 242-250.
Inouye, S. K., S. T. Bogardus, Jr., G. Vitagliano, M. M. Desai, C. S. Williams, J. N. Grady and J. D. Scinto (2003). "Burden of illness score for elderly persons: risk adjustment incorporating the cumulative impact of diseases, physiologic abnormalities, and functional impairments." Med Care **41**(1): 70-83.
Inouye, S. K., C. H. van Dyck, C. A. Alessi, S. Balkin, A. P. Siegal and R. I. Horwitz (1990). "Clarifying confusion: the confusion assessment method. A new method for detection of delirium." Ann Intern Med **113**(12): 941-948.
Inouye, S. K., R. G. Westendorp and J. S. Saczynski (2014). "Delirium in elderly people." Lancet **383**(9920): 911-922.
Jack, C. R., Jr., R. C. Petersen, Y. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, E. G. Tangalos and E. Kokmen (1998). "Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease." Neurology **51**(4): 993-999.
Jackson, T. A., D. Wilson, S. Richardson and J. M. Lord (2016). "Predicting outcome in older hospital patients with delirium: a systematic literature review." Int J Geriatr Psychiatry **31**(4): 392-399.
Jacobson, S. and H. Jerrier (2000). "EEG in delirium." Semin Clin Neuropsychiatry **5**(2): 86-92.
Jiji, S., K. A. Smitha, A. K. Gupta, V. P. Pillai and R. S. Jayasree (2013). "Segmentation and volumetric analysis of the caudate nucleus in Alzheimer's disease." Eur J Radiol **82**(9): 1525-1530.
Khachaturian, A. S., K. M. Hayden, J. W. Devlin, L. A. Fleisher, S. L. Lock, C. Cunningham, E. S. Oh, T. G. Fong, D. M. Fick, E. R. Marcantonio, V. Iyengar, K. Rockwood, G. A. Kuchel, R. G. Eckenhoff, A. M. J. MacLullich, R. N. Jones, D. Davis, P. M. D'Antonio, K. N. Fargo, M. S. Albert, J. D. Williamson, S. M. Ling, J. Weiss, J. Karlawish, R. C. Petersen, D. G. Blazer, Z. S. Khachaturian and S. K. Inouye (2020). "International drive to illuminate delirium: A developing public health blueprint for action." Alzheimers Dement **16**(5): 711-725.
Knaus, W. A., D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano and et al. (1991). "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults." Chest **100**(6): 1619-1636.
Knaus, W. A., D. P. Wagner and J. Lynn (1991). "Short-term mortality predictions for critically ill hospitalized adults: science and ethics." Science **254**(5030): 389-394.
Laurila, J. V., K. H. Pitkala, T. E. Strandberg and R. S. Tilvis (2003). "The impact of different diagnostic criteria on prevalence rates for delirium." Dement Geriatr Cogn Disord **16**(3): 156-162.

Laurila, J. V., K. H. Pitkala, T. E. Strandberg and R. S. Tilvis (2004). "Delirium among patients with and without dementia: does the diagnosis according to the DSM-IV differ from the previous classifications?" Int J Geriatr Psychiatry **19**(3): 271-277.
Lindroth, H., L. Bratzke, S. Purvis, R. Brown, M. Coburn, M. Mrkobrada, M. T. V. Chan, D. H. J. Davis, P. Pandharipande, C. M. Carlsson and R. D. Sanders (2018). "Systematic review of prediction models for delirium in the older adult inpatient." BMJ Open **8**(4): e019223.
Lindroth, H., L. Bratzke, S. Twadell, P. Rowley, J. Kildow, M. Danner, L. Turner, B. Hernandez, R. Brown and R. D. Sanders (2019). "Predicting postoperative delirium severity in older adults: The role of surgical risk and executive function." Int J Geriatr Psychiatry **34**(7): 1018-1028.
Lipowski, Z. J. (1983). "Transient cognitive disorders (delirium, acute confusional states) in the elderly." Am J Psychiatry **140**(11): 1426-1436.
Liptzin, B., S. E. Levkoff, P. D. Cleary, D. M. Pilgrim, C. H. Reilly, M. Albert and T. T. Wetle (1991). "An empirical study of diagnostic criteria for delirium." Am J Psychiatry **148**(4): 454-457.
Lopez-Rodriguez, A. B., E. Hennessy, C. L. Murray, A. Nazmi, H. J. Delaney, D. Healy, S. G. Fagan, M. Rooney, E. Stewart, A. Lewis, N. de Barra, P. Scarry, L. Riggs-Miller, D. Boche, M. O. Cunningham and C. Cunningham (2021). "Acute systemic inflammation exacerbates neuroinflammation in Alzheimer's disease: IL-1β drives amplified responses in primed astrocytes and neuronal network dysfunction." Alzheimers Dement **17**(10): 1735-1755.
Lowsky, D. J., S. J. Olshansky, J. Bhattacharya and D. P. Goldman (2014). "Heterogeneity in healthy aging." J Gerontol A Biol Sci Med Sci **69**(6): 640-649.
MacKnight, C. and K. Rockwood (1995). "A Hierarchical Assessment of Balance and Mobility." Age Ageing **24**(2): 126-130.
Maldonado, J. R. (2017). "Acute Brain Failure: Pathophysiology, Diagnosis, Management, and Sequelae of Delirium." Crit Care Clin **33**(3): 461-519.
Maldonado, J. R. (2018). "Delirium pathophysiology: An updated hypothesis of the etiology of acute brain failure." Int J Geriatr Psychiatry **33**(11): 1428-1457.
Mandelli, M. L., P. Vitali, M. Santos, M. Henry, K. Gola, L. Rosenberg, N. Dronkers, B. Miller, W. W. Seeley and M. L. Gorno-Tempini (2016). "Two insular regions are differentially involved in behavioral variant FTD and nonfluent/agrammatic variant PPA." Cortex **74**: 149-157.
Mathews, S. B., S. E. Arnold and C. N. Epperson (2014). "Hospitalization and cognitive decline: Can the nature of the relationship be deciphered?" Am J Geriatr Psychiatry **22**(5): 465-480.
McCleary, E. and P. Cumming (2015). "Improving early recognition of delirium using SQiD (Single Question to identify Delirium): a hospital based quality improvement project." BMJ Qual Improv Rep **4**(1).
Meagher, D. J., M. Leonard, S. Donnelly, M. Conroy, D. Adamis and P. T. Trzepacz (2011). "A longitudinal study of motor subtypes in delirium: relationship with other phenomenology, etiology, medication exposure and prognosis." J Psychosom Res **71**(6): 395-403.
Meagher, J., M. Leonard, L. Donoghue, N. O'Regan, S. Timmons, C. Exton, W. Cullen, C. Dunne, D. Adamis, A. J. Maclullich and D. Meagher (2015). "Months backward test: A review of its use in clinical studies." World J Psychiatry **5**(3): 305-314.
Mitasova, A., M. Kostalova, J. Bednarik, R. Michalcakova, T. Kasparek, P. Balabanova, L. Dusek, S. Vohanka and E. W. Ely (2012). "Poststroke delirium incidence and outcomes: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU)." Crit Care Med **40**(2): 484-490.
Mitchell, A. J., D. Shukla, H. A. Ajumal, B. Stubbs and T. A. Tahir (2014). "The Mini-Mental State Examination as a diagnostic and screening test for delirium: systematic review and meta-analysis." Gen Hosp Psychiatry **36**(6): 627-633.
Morandi, A., D. Davis, G. Bellelli, R. C. Arora, G. A. Caplan, B. Kamholz, A. Kolanowski, D. M. Fick, S. Kreisel, A. MacLullich, D. Meagher, K. Neufeld, P. P. Pandharipande, S. Richardson, A. J. Slooter, J. P. Taylor, C. Thomas, Z. Tieges, A. Teodorczuk, P. Voyer and J. L. Rudolph (2017). "The Diagnosis of Delirium Superimposed on Dementia: An Emerging Challenge." J Am Med Dir Assoc **18**(1): 12-18.

Morandi, A., J. McCurley, E. E. Vasilevskis, D. M. Fick, G. Bellelli, P. Lee, J. C. Jackson, S. D. Shenkin, Marcotrabucchi, J. Schnelle, S. K. Inouye, E. W. Ely and A. MacLullich (2012). "Tools to detect delirium superimposed on dementia: a systematic review." J Am Geriatr Soc **60**(11): 2005-2013.
Morandi, A., B. P. Rogers, M. L. Gunther, K. Merkle, P. Pandharipande, T. D. Girard, J. C. Jackson, J. Thompson, A. K. Shintani, S. Geevarghese, R. R. Miller, 3rd, A. Canonico, C. J. Cannistraci, J. C. Gore, E. W. Ely and R. O. Hopkins (2012). "The relationship between delirium duration, white matter integrity, and cognitive impairment in intensive care unit survivors as determined by diffusion tensor imaging: the VISIONS prospective cohort magnetic resonance imaging study*." Crit Care Med **40**(7): 2182-2189.
Mormino, E. C., K. V. Papp, D. M. Rentz, M. C. Donohue, R. Amariglio, Y. T. Quiroz, J. Chhatwal, G. A. Marshall, N. Donovan, J. Jackson, J. R. Gatchel, B. J. Hanseeuw, A. P. Schultz, P. S. Aisen, K. A. Johnson and R. A. Sperling (2017). "Early and late change on the preclinical Alzheimer's cognitive composite in clinically normal older individuals with elevated amyloid β." Alzheimers Dement **13**(9): 1004-1012.
Naksuk, N., C. Thongprayoon, J. Y. Park, S. Sharma, P. Gaba, A. N. Rosenbaum, T. Peeraphatdit, T. Y. Hu, M. R. Bell, V. Herasevich, P. A. Brady, S. Kapa and S. J. Asirvatham (2017). "Editor's Choice-Clinical impact of delirium and antipsychotic therapy: 10-Year experience from a referral coronary care unit." Eur Heart J Acute Cardiovasc Care **6**(6): 560-568.
Newman, M. F., H. P. Grocott, J. P. Mathew, W. D. White, K. Landolfo, J. G. Reves, D. T. Laskowitz, D. B. Mark and J. A. Blumenthal (2001). "Report of the substudy assessing the impact of neurocognitive function on quality of life 5 years after cardiac surgery." Stroke **32**(12): 2874-2881.
Ng, K. T., C. J. Shubash and J. S. Chong (2019). "The effect of dexmedetomidine on delirium and agitation in patients in intensive care: systematic review and meta-analysis with trial sequential analysis." Anaesthesia **74**(3): 380-392.
Nikooie, R., K. J. Neufeld, E. S. Oh, L. M. Wilson, A. Zhang, K. A. Robinson and D. M. Needham (2019). "Antipsychotics for Treating Delirium in Hospitalized Adults: A Systematic Review." Ann Intern Med.
Numan, T., M. van den Boogaard, A. M. Kamper, P. J. T. Rood, L. M. Peelen and A. J. C. Slooter (2019). "Delirium detection using relative delta power based on 1-minute single-channel EEG: a multicentre study." Br J Anaesth **122**(1): 60-68.
Oskolkov, N. (2019). "How to tune hyperparameters of tSNE." Towards Data Science.
Pandharipande, P. P., T. D. Girard, J. C. Jackson, A. Morandi, J. L. Thompson, B. T. Pun, N. E. Brummel, C. G. Hughes, E. E. Vasilevskis, A. K. Shintani, K. G. Moons, S. K. Geevarghese, A. Canonico, R. O. Hopkins, G. R. Bernard, R. S. Dittus and E. W. Ely (2013). "Long-term cognitive impairment after critical illness." N Engl J Med **369**(14): 1306-1316.
Partridge, J. S., F. C. Martin, D. Harari and J. K. Dhesi (2013). "The delirium experience: what is the effect on patients, relatives and staff and what can be done to modify this?" Int J Geriatr Psychiatry **28**(8): 804-812.
Petersen, S. E. and M. I. Posner (2012). "The attention system of the human brain: 20 years after." Annu Rev Neurosci **35**: 73-89.
Pezzullo, L., J. Streatfeild, J. Hickson, A. Teodorczuk, M. R. Agar and G. A. Caplan (2019). "Economic impact of delirium in Australia: a cost of illness study." BMJ Open **9**(9): e027514.
Pitkala, K. H., J. V. Laurila, T. E. Strandberg and R. S. Tilvis (2005). "Prognostic significance of delirium in frail older people." Dement Geriatr Cogn Disord **19**(2-3): 158-163.
Powers, B. W., S. K. Chaguturu and T. G. Ferris (2015). "Optimizing high-risk care management." Jama **313**(8): 795-796.
Roberts, S. E. and M. J. Goldacre (2003). "Time trends and demography of mortality after fractured neck of femur in an English population, 1968-98: database study." Bmj **327**(7418): 771-775.
Rockwood, K., J. Goodman, M. Flynn and P. Stolee (1996). "Cross-validation of the Delirium Rating Scale in older patients." J Am Geriatr Soc **44**(7): 839-842.
Rohatgi, N., Y. Weng, J. Bentley, M. G. Lansberg, J. Shepard, D. Mazur, N. Ahuja and J. Hopkins (2019). "Initiative for Prevention and Early Identification of Delirium in Medical-Surgical Units: Lessons Learned in the Past Five Years." Am J Med **132**(12): 1421-1430.e1428.
Rutter, L. M., E. Nouzova, D. J. Stott, C. J. Weir, V. Assi, J. H. Barnett, C. Clarke, N. Duncan, J. Evans, S. Green, K. Hendry, M. McGinlay, J. McKeever, D. G. Middleton, S. Parks, R. Shaw, E. Tang, T. Walsh, A. J.

Weir, E. Wilson, T. Quasim, A. M. J. MacLullich and Z. Tieges (2018). "Diagnostic test accuracy of a novel smartphone application for the assessment of attention deficits in delirium in older hospitalised patients: a prospective cohort study protocol." BMC Geriatr **18**(1): 217.
Saczynski, J. S., E. R. Marcantonio, L. Quach, T. G. Fong, A. Gross, S. K. Inouye and R. N. Jones (2012). "Cognitive trajectories after postoperative delirium." N Engl J Med **367**(1): 30-39.
Schuurmans, M. J., L. M. Shortridge-Baggett and S. A. Duursma (2003). "The Delirium Observation Screening Scale: a screening instrument for delirium." Res Theory Nurs Pract **17**(1): 31-50.
Searle, S. D., A. Mitnitski, E. A. Gahbauer, T. M. Gill and K. Rockwood (2008). "A standard procedure for creating a frailty index." BMC Geriatr **8**: 24.
Sepulveda, E., J. G. Franco, P. T. Trzepacz, A. M. Gaviria, E. Vinuelas, J. Palma, G. Ferre, I. Grau and E. Vilella (2015). "Performance of the Delirium Rating Scale-Revised-98 Against Different Delirium Diagnostic Criteria in a Population With a High Prevalence of Dementia." Psychosomatics **56**(5): 530-541.
SIGN (2019). "Risk reduction and management of delirium."
Skelly, D. T., E. W. Griffin, C. L. Murray, S. Harney, C. O'Boyle, E. Hennessy, M. A. Dansereau, A. Nazmi, L. Tortorelli, J. N. Rawlins, D. M. Bannerman and C. Cunningham (2019). "Acute transient cognitive dysfunction and acute brain injury induced by systemic inflammation occur by dissociable IL-1-dependent mechanisms." Mol Psychiatry **24**(10): 1533-1548.
Teno, J. M., F. E. Harrell, Jr., W. Knaus, R. S. Phillips, A. W. Wu, A. Connors, Jr., N. S. Wenger, D. Wagner, A. Galanos, N. A. Desbiens and J. Lynn (2000). "Prediction of survival for older hospitalized patients: the HELP survival model. Hospitalized Elderly Longitudinal Project." J Am Geriatr Soc **48**(S1): S16-24.
Tieges, Z., T. Quinn, L. MacKenzie, D. Davis, G. Muniz-Terrera, A. M. J. MacLullich and S. D. Shenkin (2021). "Association between components of the delirium syndrome and outcomes in hospitalised adults: a systematic review and meta-analysis." BMC Geriatr **21**(1): 162.
Todd, A., S. Blackley, J. K. Burton, D. J. Stott, E. W. Ely, Z. Tieges, A. M. J. MacLullich and S. D. Shenkin (2017). "Reduced level of arousal and increased mortality in adult acute medical admissions: a systematic review and meta-analysis." BMC Geriatr **17**(1): 283.
Tsui, A., D. Kuh, M. Richards and D. Davis (2018). "Delirium symptoms are associated with decline in cognitive function between ages 53 and 69 years: Findings from a British birth cohort study." Alzheimers Dement **14**(5): 617-622.
Tsui, A., S. D. Searle, H. Bowden, K. Hoffmann, J. Hornby, A. Goslett, M. Weston-Clarke, L. H. Howes, R. Street, R. Perera, K. Taee, C. Kustermann, P. Chitalu, B. Razavi, F. Magni, D. Das, S. Kim, N. Chaturvedi, E. L. Sampson, K. Rockwood, C. Cunningham, E. W. Ely, S. J. Richardson, C. Brayne, G. M. Terrera, Z. Tieges, A. MacLullich and D. Davis (2022). "The effect of baseline cognition and delirium on long-term cognitive impairment and mortality: a prospective population-based study." Lancet Healthy Longev **3**(4): e232-e241.
Tsui, A., P. D. Tudosiu, M. Brudfors, A. Jha, J. Cardoso, S. Ourselin, J. Ashburner, G. Rees, D. Davis and P. Nachev (2023). "Predicting mortality in acutely hospitalised older patients: the impact of model dimensionality." BMC Med **21**(1): 10.
Tsui, A., N. Yeo, S. D. Searle, H. Bowden, K. Hoffmann, J. Hornby, A. Goslett, M. Weston-Clarke, D. Lanham, P. Hogan, A. Seeley, M. Rawle, N. Chaturvedi, E. L. Sampson, K. Rockwood, C. Cunningham, E. W. Ely, S. J. Richardson, C. Brayne, G. M. Terrera, Z. Tieges, A. M. J. MacLullich and D. Davis (2023). "Extremes of baseline cognitive function determine the severity of delirium: a population study." Brain **146**(5): 2132-2141.
van der Kall, L. M., T. Truong, S. C. Burnham, V. Doré, R. S. Mulligan, S. Bozinovski, F. Lamb, P. Bourgeat, J. Fripp, S. Schultz, Y. Y. Lim, S. M. Laws, D. Ames, C. Fowler, S. R. Rainey-Smith, R. N. Martins, O. Salvado, J. Robertson, P. Maruff, C. L. Masters, V. L. Villemagne and C. C. Rowe (2021). "Association of β-Amyloid Level, Clinical Progression, and Longitudinal Cognitive Change in Normal Older Individuals." Neurology **96**(5): e662-e670.
van der Maaten, L. and G. Hinton (2008). "Visualising data using t-SNE." Journal of machine learning research **9**: 2579-2605.

van Montfort, S. J. T., E. van Dellen, C. J. Stam, A. H. Ahmad, L. J. Mentink, C. W. Kraan, A. Zalesky and A. J. C. Slooter (2019). "Brain network disintegration as a final common pathway for delirium: a systematic review and qualitative meta-analysis." Neuroimage Clin **23**: 101809.
van Velthuijsen, E. L., S. M. Zwakhalen, R. M. Warnier, W. J. Mulder, F. R. Verhey and G. I. Kempen (2016). "Psychometric properties and feasibility of instruments for the detection of delirium in older hospitalized patients: a systematic review." Int J Geriatr Psychiatry **31**(9): 974-989.
van Walraven, C. (2014). "The Hospital-patient One-year Mortality Risk score accurately predicted long-term death risk in hospitalized patients." J Clin Epidemiol **67**(9): 1025-1034.
van Walraven, C. and A. J. Forster (2017). "The HOMR-Now! Model Accurately Predicts 1-Year Death Risk for Hospitalized Patients on Admission." Am J Med **130**(8): 991.e999-991.e916.
van Walraven, C., F. A. McAlister, J. A. Bakal, S. Hawken and J. Donze (2015). "External validation of the Hospital-patient One-year Mortality Risk (HOMR) model for predicting death within 1 year after hospital admission." Cmaj **187**(10): 725-733.
Voyer, P., N. Champoux, J. Desrosiers, P. Landreville, J. McCusker, J. Monette, M. Savoie, S. Richard and P. H. Carmichael (2015). "Recognizing acute delirium as part of your routine [RADAR]: a validation study." BMC Nurs **14**: 19.
Wallace, L. M. K., O. Theou, J. Godin, M. J. Andrew, D. A. Bennett and K. Rockwood (2019). "Investigation of frailty as a moderator of the relationship between neuropathology and dementia in Alzheimer's disease: a cross-sectional analysis of data from the Rush Memory and Aging Project." Lancet Neurol **18**: 177-184.
Wallace, L. M. K., O. Theou, J. Godin, M. K. Andrew, D. A. Bennett and K. Rockwood (2019). "Investigation of frailty as a moderator of the relationship between neuropathology and dementia in Alzheimer's disease: a cross-sectional analysis of data from the Rush Memory and Aging Project." Lancet Neurol **18**(2): 177-184.
Wassenaar, A., L. Schoonhoven, J. W. Devlin, F. M. P. van Haren, A. J. C. Slooter, P. G. Jorens, M. van der Jagt, K. S. Simons, I. Egerod, L. D. Burry, A. Beishuizen, J. Matos, A. R. T. Donders, P. Pickkers and M. van den Boogaard (2019). "External Validation of Two Models to Predict Delirium in Critically Ill Adults Using Either the Confusion Assessment Method-ICU or the Intensive Care Delirium Screening Checklist for Delirium Assessment." Crit Care Med **47**(10): e827-e835.
Whittamore, K. H., S. E. Goldberg, J. R. Gladman, L. E. Bradshaw, R. G. Jones and R. H. Harwood (2014). "The diagnosis, prevalence and outcome of delirium in a cohort of older people with mental health problems on general hospital wards." Int J Geriatr Psychiatry **29**(1): 32-40.
Witlox, J., L. S. Eurelings, J. F. de Jonghe, K. J. Kalisvaart, P. Eikelenboom and W. A. van Gool (2010). "Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis." Jama **304**(4): 443-451.
Wong, E. K., J. Y. Lee, A. S. Surendran, K. Nair, N. Della Maestra, M. Migliarini, J. A. St Onge and C. J. Patterson (2018). "Nursing perspectives on the confusion assessment method: a qualitative focus group study." Age Ageing **47**(6): 880-886.
Wu, Y., Z. Shi, M. Wang, Y. Zhu, C. Li, G. Li, E. R. Marcantonio, Z. Xie and Y. Shen (2015). "Different MMSE Score Is Associated with Postoperative Delirium in Young-Old and Old-Old Adults." PLoS One **10**(10): e0139879.
Yourman, L. C., S. J. Lee, M. A. Schonberg, E. W. Widera and A. K. Smith (2012). "Prognostic indices for older adults: a systematic review." Jama **307**(2): 182-192.
Zucchelli, A., D. L. Vetrano, G. Grande, A. Calderon-Larranaga, L. Fratiglioni, A. Marengoni and D. Rizzuto (2019). "Comparing the prognostic value of geriatric health indicators: a population-based study." BMC Med **17**(1): 185.

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

   a) **What is the title of the manuscript?**

   Extremes of baseline cognitive function determine the severity of delirium: a population study

   b) **Please include a link to or doi for the work**

   https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad062/7059705

   c) **Where was the work published?**

   Brain

   d) **Who published the work?** (e.g. OUP)

   Oxford University Press

   e) **When was the work published?**

   28th February 2023

   f) **List the manuscript's authors in the order they appear on the publication**

Alex Tsui, Natalie Yeo, Samuel D Searle, Helen Bowden, Katrin Hoffmann, Joanne Hornby, Arley Goslett, Maryse Weston-Clarke, David Lanham, Patrick Hogan, Anna Seeley, Mark Rawle, Nish Chaturvedi, Elizabeth L Sampson, Kenneth Rockwood, Colm Cunningham, E Wesley Ely, Sarah J Richardson, Carol Brayne, Graciela Muniz Terrera, Zoë Tieges, Alasdair M J MacLullich, Daniel Davis

   g) **Was the work peer reviewed?**

   Yes

   h) **Have you retained the copyright?**

   Yes

   i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

   No

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☒

   *I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   a) **What is the current title of the manuscript?**

   Click or tap here to enter text.

   b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

   Click or tap here to enter text.

   c) **Where is the work intended to be published?** (e.g. journal names)

   Click or tap here to enter text.

   d) **List the manuscript's authors in the intended authorship order**

   Click or tap here to enter text.

   e) **Stage of publication** (e.g. in submission)

   Click or tap here to enter text.

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

AT, NY, SDS, DD curated the data. AT, NY, GMT, and DDav did the formal analysis. NC, CB, KR, AM, AT and DDav acquired funding. HB, KH, AG, MWC, DL, PH, AS, MR collected the data. AT, NY, SDS, KR, CC, GMT, DD did the investigation. ELS, KR, SJR, GMT, CB, EE, CC, AM and ZT contributed to the methods. JH and KT did the project administration. AT, NY and DD wrote the original draft.

4. **In which chapter(s) of your thesis can this material be found?**

   5

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

[redacted]

Alex Tsui
Date:
8th March 2023

Supervisor/ Senior Author (where appropriate)

[redacted]

Daniel Davis
8th March 2023

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

   a) **What is the title of the manuscript?**

   The effect of baseline cognition and delirium on long-term cognitive impairment and mortality: a prospective population-based study

   b) **Please include a link to or doi for the work**

   https://doi.org/10.1016/S2666-7568(22)00013-7

   c) **Where was the work published?**

   Lancet Healthy Longetvity

   d) **Who published the work?** (e.g. OUP)

   Elsevier

   e) **When was the work published?**

   15th March 2022

   f) **List the manuscript's authors in the order they appear on the publication**

   Alex Tsui, Samuel D Searle, Helen Bowden, Katrin Hoffmann, Joanne Hornby, Arley Goslett, Maryse Weston-Clarke, Lee Hamill Howes,Rebecca Street, Rachel Perera, Kayvon Taee, Christoph Kustermann, Petronella Chitalu, Benjamin Razavi, Francesco Magni, Devajit Das, Sung Kim, Nish Chaturvedi, Elizabeth L Sampson, Kenneth Rockwood, Colm Cunningham, E Wesley Ely, Sarah J Richardson, Carol Brayne,
   Graciela Muniz Terrera, Zoë Tieges, Alasdair MacLullich, Daniel Davis

   g) **Was the work peer reviewed?**

   Yes

   h) **Have you retained the copyright?**

   Yes

   i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

   No

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☒

   *I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   f) **What is the current title of the manuscript?**

   Click or tap here to enter text.

   g) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

   Click or tap here to enter text.

   h) **Where is the work intended to be published?** (e.g. journal names)

   Click or tap here to enter text.

   i) **List the manuscript's authors in the intended authorship order**

Click or tap here to enter text.

j) **Stage of publication** (e.g. in submission)

Click or tap here to enter text.

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

AT, SDS, HB, KH, JH, AG, MW-C, LHH, RS, RP, and KT curated the data. AT, SDS, GMT, and DDav did the formal analysis. NC, CB, KR, AM, AT and DDav acquired funding. AT, SDS, HB, KH, JH, AG, MW-C, LHH, RS, RP, KT, CK, PC, BR, FM, DDas, and SK did the investigation. ELS, KR, SJR, GMT, and ZT contributed to the methods. JH and KT did the project administration. AT and DDav wrote the original draft.

4. **In which chapter(s) of your thesis can this material be found?**

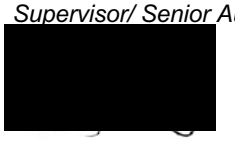
5

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Alex Tsui
Date:
8th March 2023

Supervisor/ Senior Author (where appropriate)



Daniel Davis
8th March 2023

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

   a) **What is the title of the manuscript?**

   Predicting mortality in acutely hospitalised older patients

   b) **Please include a link to or doi for the work**

   https://doi.org/10.1186/s12916-022-02698-2

   c) **Where was the work published?**

   BMC Medicine

   d) **Who published the work?** (e.g. OUP)

   Biomed Central

   e) **When was the work published?**

   8th January 2023

   f) **List the manuscript's authors in the order they appear on the publication**

   Tsui A, Tudosiu P-D, Brudfors M, Jha A, Cardoso J, Ourselin S, Ashburner J, Rees G, Davis D, Nachev P

   g) **Was the work peer reviewed?**

   Yes

   h) **Have you retained the copyright?**

   Yes

   i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

   No

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☒

   *I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   k) **What is the current title of the manuscript?**

   Click or tap here to enter text.

   l) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

   Click or tap here to enter text.

   m) **Where is the work intended to be published?** (e.g. journal names)

   Click or tap here to enter text.

   n) **List the manuscript's authors in the intended authorship order**

   Click or tap here to enter text.

   o) **Stage of publication** (e.g. in submission)

   Click or tap here to enter text.

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

AT, PDT, DD, and PN contributed to the conceptualisation of the study. AT, PDT, MB, and PN curated the data. AT, PDT, MB, and PN performed the formal analysis. AT, DD, JC, SO, GR, and PN acquired funding. AT, PDT, MB, DD, and PN performed the investigation. AT, PDT, MB, AJ, JC, SO, JA, GR, DD, and PN con- tributed to the methods. AT and PDT did the project administration. DD and PN supervised the study. AT, PDT, and PN validated the results. AT, PDT, and PN produced visualisation. AT, PDT, and PN wrote the original draft. AT, PDT, MB, AJ, JC, SO, JA, GR, DD, and PN reviewed and edited the manuscript. All authors read and approved the fnal manuscript.

4. **In which chapter(s) of your thesis can this material be found?**

   6

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

[redacted]

Alex Tsui
Date:
8th March 2023

Supervisor/ Senior Author (where appropriate)

[redacted]

Daniel Davis
8th March 2023