Outcomes adjudication **Start date: 1 May 2015.**				**Completion date: 31 March 2020**
Overall work package objectives: To develop methods for dementia case identification at scale using the UKB cohort and produce an online system for remote classification. It will validate these methods using UKB and other DPUK cohorts. 1. Provide up-to-date assessments of the accuracy of routine data sources for dementia, PD and motor neurone disease (MND) outcomes 2. Develop algorithms based on routine data sources for ascertaining, confirming and sub-classifying these outcomes 3. Develop protocols for remote confirmation and sub-classification using an online system 4. Apply these protocols for adjudication (i.e. confirmation according to standard criteria) the dementia, PD and MND outcomes occurring in UKB and other cohorts.			**Dependencies to and from other work packages, networks and themes** None that are critical although we interact with and benefit from other WPs and themes within DPUK.	
Deliverables	**Milestones**	**Milestone deadline**	**Work package dependencies**	**Person(s) responsible**
Objective 1:				
D1.1 Produce reports on accuracy of routine data sources	M1.1.1 Identification of reference standard data sources for routine data	M1.1.1 Complete	None	Cathie Sudlow
D1.2 Produce updated guidance for researchers	M1.2.1 Plan in place for comparison of routine data with reference standards	M1.2.1 Complete		
Objective 2:				
D2.1 Protocols and system complete	M2.1.1 Cross reference other relevant data sets	M2.1.1 Complete	None	Cathie Sudlow
	M2.1.2 Development and testing of automated algorithms	M2.1.2 Complete		
Objective 3:				
D3.1 Identification of additional information available in full electronic medical records	M3.1.1 Additional information identified	M3.1.1 Dec 2020	None	Cathie Sudlow

Key updates on delivery against milestones since last report

Fully met

We have published systematic reviews summarising the existing literature on the accuracy of using routinely-collected, coded datasets to identify dementia, MND and PD

Fully met

Algorithms have been developed for dementia, MND and PD

Partially met/in progress

Phase 1 protocols (ICD9/10 and self-report data) have been completed for dementia, MND and PD

Phase 2 protocols (ICD 9/10, self-report data and READ v2 and 3) for these conditions have been prepared and aim to be released in 2019.

Phase 3 protocols (accessing medical record free text) for these conditions have been done in a limited way in NHS Lothian, we are looking to explore Natural Language processing options in NHS Lothian.

Fully met

The Phase 1 algorithms were added to UK Biobank cohort in March 2019 (http://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=42) with the phase 2 algorithms released to researchers in September 2019 with the addition of the GP data to the cohort.
Applying the algorithms to other cohorts: detailed documentation prepared and returned to DUK to enable other population-based cohorts to apply these. Looking to apply these algorithms to the Generation Scotland cohort by March 2020.

Summary of plan to deliver on outstanding work

- Plan to apply the developed algorithms to Generation Scotland – March 2020
- Plan to explore Natural Language processing in NHS Lothian – March 2020

Risks	**Mitigation**
1) n/a	1)

Team members (most with part time commitment to DPUK and part to other projects, including UK Biobank)
Cathie Sudlow (PI), Emma Davidson (clinical research fellow), Robin Flaig (project manager), Clifford Nangle (data analyst), Christian Schnier (epidemiologist), Mary Morrissey (research nurse coordinator), Aidan Hutchison (data manager and programmer)

Team members involved with the project but not funded by DPUK
Tim Wilkinson (PhD student, previously a DPUK clinical research fellow)

Outcomes

Publication output

1. **Horrocks S, Wilkinson T, Schnier C, Ly A, Woodfield R, Rannikmäe K, Quinn TJ, Sudlow CLM. (2017) Accuracy of routinely-collected healthcare data for identifying motor neurone disease cases: A systematic review. PLoS ONE 12(2): e0172639. https://doi.org/10.1371/journal.pone.0172639**

In this novel systematic review, we summarised evidence from validation studies that used routine datasets to identify motor neurone disease cases, allowing researchers around the world who are using these datasets to understand their accuracy. This work also identified gaps in knowledge, indicating where further validation research is required.

2. **Wilkinson T, Ly A, Schnier C, Rannikmäe K, Bush K, Brayne C, Quinn TJ, Sudlow CLM. Identifying dementia cases with routinely-collected health data: a systematic review. Alzheimers Dement. 2018 Aug;14(8):1038–51.**

We systematically reviewed the international literature on the accuracy of using routine datasets to identify dementia cases - a previously unaddressed area of evidence synthesis. We identified high accuracy datasets suitable for dementia research and highlighted areas requiring further research, advising cohorts to conduct validation studies where such gaps exist.

3. **Harding Z, Wilkinson T, Stevenson A, Horrocks S, Ly A, Schnier C, Breen D, Rannikmäe K, Sudlow CLM. Accuracy of routinely-collected health data for identifying Parkinson's disease and parkinsonism cases: a systematic review. PLOS ONE. 2019 Jan 31;14(1):e0198736.**

We systematically reviewed studies validating the accuracy of using routinely-collected healthcare datasets to identify Parkinson's disease cases. We found relatively wide-ranging accuracy estimates and identified gaps in current knowledge. We highlighted areas where further, context-specific validation studies are warranted and identified additional datasets that may assist future algorithm development.

4. **Wilkinson T, Schnier C, Bush K, Rannikmäe K, Henshall D, Lerpiniere C, Allen NE, Flaig R, Russ T, Bathgate D, Suvankar P, O'Brien JT, Sudlow CLM. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. Eur J Epidemiol. 2019 Feb 26. doi: 10.1007/s10654-019-00499-1.**

Using UK Biobank (UKB) data, we conducted a validation study of using routinely-collected healthcare datasets (primary care, hospital admissions and mortality data) to identify dementia cases. We identified three algorithms that are likely to be highly predictive of all-cause dementia. UKB will use these to generate derived dementia outcome variables for researchers.

5. **Wilkinson T, Schnier C et al. Cohort Data resource profile: The Secure Anonymised Information Linkage databank Dementia e-cohort (SAIL-DeC). Being resubmitted at International Journal of Population Health Sciences.**

We aimed to apply reproducible methods to create the SAIL dementia e-cohort (SAIL-DeC). We created SAIL-DeC with a view to maximising its utility for a broad range of research questions whilst minimising duplication of effort for researchers. From 4.4 million unique participants in SAIL, 1.2 million met the cohort inclusion criteria, resulting in 18.8 million person-years of follow-up. Among the dementia cases, the median duration of observation time was 14 years.

Project narrative

This work has investigated the accuracy of using routinely-collected primary care, hospital admissions and mortality datasets to identify dementia, motor neurone disease (MND) or Parkinson's disease (PD) cases. Systematic reviews on the accuracy of these datasets for dementia (Wilkinson et al. 2018), MND (Horrocks et al. 2017) and PD (Harding et al. 2019) were published. These reviews summarised the findings of existing validation studies, and identified gaps in current knowledge, particularly with respect to a UK setting. Following on from this, a validation study of UK primary care, hospital admissions and mortality datasets was conducted using data from UK Biobank (Wilkinson et al. 2019). In this study, coded diagnoses from routine datasets were compared to the full text medical record, leading to the identification of algorithms that can be employed by UK cohorts to identify participants who develop dementia during follow up. UK Biobank has used these findings to provide derived dementia outcome variables for researchers, thereby facilitating dementia research using the resource. These algorithms were made freely available to the research community, enabling other UK cohorts to adopt this approach to dementia case identification. There was a plan to scale up the adjudication programme up to include Leeds and Birmingham, however this required working with UK Biobank who has decided not to support this extension of validation work at this time and so it is unable to be completed.

The application of these algorithms to Generation Scotland is going to be undertaken in the next few months. In addition, the team will investigate the possibility of running NLP and machine learning techniques on full hospital patient records in NHS Lothian.

A separate project from the same group, with collaborators at Swansea University, has used routinely-collected healthcare data to create an electronic cohort (e-cohort) for dementia research. Utilising the SAIL Databank (https://saildatabank.com/), this e-cohort contains data for 1.2 million people, of whom 130,000 developed dementia during follow-up. Access to the e-cohort is facilitated through DPUK.