Clinically reported covert cerebrovascular disease and risk of neurological disease: a whole-population cohort of 395,273 people using natural language processing

Matthew H Iveson PhD [1] ORCID: 0000-0002-7242-0456

Mome Mukerjee MPhil [2,3] ORCID: 0000-0002-3083-436X

Emma M Davidson MD [1] ORCID: 0000-0002-5182-8495

Huayu Zhang PhD [2] ORCID: 0000-0003-1625-4067

Laura Sherlock MSc [1] ORCID: 0000-0002-1732-9713

Emily L Ball PhD [1] ORCID: 0000-0002-7445-9581

Grant Mair MD [1] ORCID: 0000-0003-2189-443X

Alice Hosking MRCP [1] ORCID: 0000-0001-6900-7778

Prof. Heather Whalley PhD [1] ORCID: 0000-0002-4505-8869

Michael T C Poon PhD [1] ORCID: 0000-0002-0053-2184

Prof. Joanna M Wardlaw MD, FRCR, FMedSci [1,4] ORCID: 0000-0002-9812-6642

Prof. David Kent PhD [5] ORCID: 0000-0002-9205-5070

Richard Tobin MA [6] ORCID: 0000-0003-4050-7888

Claire Grover PhD [6] ORCID: 0000-0001-7019-1250

Beatrice Alex PhD [7,8,9] ORCID: 0000-0002-7279-1476

Prof. William N Whiteley FRCP [1,2,10] ORCID: 0000-0002-4816-8991

1. Centre for Clinical Brain Sciences, University of Edinburgh
2. Centre for Clinical Informatics, Usher Institute, University of Edinburgh
3. Nuffield Department of Primary Care Health Sciences, University of Oxford
4. UK Dementia Research Institute, University of Edinburgh
5. Graduate School of Biomedical Sciences, Tufts University
6. School of Informatics, University of Edinburgh
7. School of Literatures, Languages and Cultures, University of Edinburgh
8. Edinburgh Futures Institute, University of Edinburgh
9. Advanced Care Research Centre, University of Edinburgh
10. British Heart Foundation Data Science Centre, HDRUK, London

Corresponding author: Matthew H. Iveson, Centre for Clinical Brain Sciences, The University of Edinburgh, Chancellor's Building, Edinburgh EH16 4SB, UK. Tel: (+44) 0131 650 4005. E-mail: Matthew.Iveson@ed.ac.uk

Abstract

Background

The relevance of covert cerebrovascular disease (CCD) in practice is uncertain, in part because estimation of risk in whole clinical populations is difficult. To address this gap, we measured clinically reported CCD in a large clinical cohort using natural language processing (NLP) and estimated subsequent disease risk in linked health record data.

Methods

From all people with brain imaging in Scotland from 2010 to 2018, we selected people with no prior hospitalisation for neurological disease. Four phenotypes were identified with NLP of imaging reports: white matter hypoattenuation or hyperintensities (WMH), lacunes, cortical infarcts and cerebral atrophy. Hazard ratios (aHR) for stroke, dementia, and Parkinson's disease (conditions previously associated with CCD), epilepsy (a brain-based control condition) and colorectal cancer (a non-brain control condition), adjusted for age, sex, deprivation, region, scan modality, and pre-scan healthcare, were calculated for each phenotype.

Findings

From 395,273 people with brain imaging and no history of neurological disease, 145,978 (37%) had ≥1 phenotype. For each phenotype, the aHR of any stroke was: WMH 1·4 (95%CI: 1·3–1·4), lacunes 1·6 (1·5–1·6), cortical infarct 1·7 (1·6–1·8), and cerebral atrophy 1·1 (1·0–1·1). The aHR of any dementia was: WMH, 1·3 (1·3–1·3), lacunes, 1·0 (0·9–1·0), cortical infarct 1·1 (1·0–1·1) and cerebral atrophy 1·7 (1·7–1·7). The aHR of Parkinson's disease was, in people with a report of: WMH 1·1 (1·0–1·2), lacunes 1·1 (0·9–1·2), cortical infarct 0·7 (0·6–0·9) and cerebral atrophy 1·4 (1·3–1·5). The aHRs between CCD phenotypes and epilepsy and colorectal cancer overlapped the null.

Interpretation

NLP identified CCD and atrophy phenotypes from routine clinical image reports, and these had important associations with future stroke, dementia and Parkinson's disease. Prevention of neurological disease in people with CCD should be a priority for healthcare providers and policymakers.

Funding

The Chief Scientist's Office, the Medical Research Council, the Alzheimer's Society, Health Data Research UK, the Wellcome Trust, Research Data Scotland, MQ – Transforming Mental Health, The Alan Turing Institute, the National Institute for Health Research, and the Stroke Association.

Keywords

Stroke, dementia, cerebrovascular disease, radiology reports, natural language processing

Research in context

Evidence before this study

Systematic reviews of magnetic resonance imaging (MRI) in cohort studies show that people with asymptomatic, covert cerebrovascular disease (CCD) have an increased risk of stroke, dementia and Parkinson's disease. Covert brain infarcts, identified with natural language processing (NLP) show similar associations in one US-based insurance system. However, there is a lack of data on clinically-reported CCD from whole populations.

Added value of this study

This study used a validated NLP algorithm to identify CCD and cerebral atrophy from both MRI and computed tomography (CT) imaging reports generated during routine healthcare in >395K people in Scotland. It also distinguished between three CCD phenotypes – white matter hypoattenuation/hyperintensities, lacunes, cortical infarcts – and cerebral atrophy, and their associations with stroke and dementia and their subtypes. In adjusted models, we demonstrate higher risk of dementia (particularly Alzheimer's disease) in people with atrophy, and higher risk of stroke in people with cortical infarcts. However, associations with an age-associated control outcome (colorectal cancer) were neutral, supporting a causal relationship. It also highlights differential associations between cerebral atrophy and dementia and cortical infarcts and stroke risk.

Implications of all the available evidence

CCD or atrophy on brain imaging reports in routine clinical practice is associated with a higher risk of stroke or dementia. Evidence is needed to support treatment strategies to reduce this risk. NLP can identify these important, otherwise uncoded, disease phenotypes, allowing research at scale into imaging-based biomarkers of dementia and stroke.

Introduction

Covert cerebrovascular disease (CCD) is a common incidental finding after brain imaging. CCD imaging phenotypes include subcortical and periventricular white matter hypoattenuation and hyperintensities (WMH), small subcortical (lacunar) infarcts and large artery atherothrombo- or cardio-embolic (cortical) infarcts in people who have never reported symptoms of stroke or transient ischaemic attack (TIA). Systematic reviews of cohorts screened by magnetic resonance imaging (MRI) show that CCD prevalence is much higher than symptomatic stroke,[1–3] and that findings of CCD are independently associated with stroke and dementia in population-based cohorts[4]. However, these findings may not generalise to clinical practice, because participants in cohort studies are not representative of people seeking healthcare, [5] and because neuroimaging reports generated during research are more standardised than clinical reports. Furthermore, while most studies identify CCD with MRI, computed tomography (CT) is more commonly used in clinical practice, particularly for those with minor symptoms for whom there is clinical uncertainty. Therefore, the clinical relevance of clinically reported CCD is often uncertain.

CCD does not have a diagnostic code in common classifications (e.g., International Classification of Diseases (ICD)-11), hence it is difficult to study in populations representative of clinical practice. Natural language processing (NLP) can help to identify hard to find phenotypes in medical notes and clinical reports. We have developed and validated a rules-based NLP algorithm (EdIE-R) with excellent recall and precision to identify brain imaging phenotypes from brain imaging radiology reports.[6] This algorithm performed with high accuracy, compared with a reference standard of expert reads of brain imaging reports or brain images.[6–9] Therefore NLP could reveal CCD and atrophy phenotypes in whole populations of patients who have had brain imaging during routine healthcare.

Here, we used NLP to identify clinically reported CCD and cerebral atrophy on CT and MRI brain imaging reports from routine healthcare in the Scottish population for up to 12 years after index imaging. We estimated the absolute risks and association of clinically reported CCD and atrophy with incident stroke, dementia, and Parkinson's disease, and with epilepsy (a brain-based control condition not directly related to CCD) and colorectal cancer (a non-brain control condition).

Methods

Study design and participants

We used de-identified individual-level data, accessed through the National Health Service (NHS) Scotland's National Safe Haven provided by Public Health Scotland (PHS). PHS deterministically linked general hospital inpatient records (Scottish morbidity record (SMR) 01), mental health inpatient records (SMR 04), community prescribing (Prescribing Information System (PIS)), cancer registrations (SMR 06), NRS death records and brain imaging reports from Scottish Medical Imaging with the community health index (CHI) number, a unique patient identifier, between 1st January 2008 and 31st December 2020.[10]

We extracted data from the whole population of Scotland with at least one NHS Scotland hospitalisation, outpatient or prescribing record (5,667,954 people). We selected all 785,331 people with a brain MRI or CT scan between January 2010 and August 2018. We excluded people with missing demographic data, an imaging date that conflicted with emigration or death, reports from research studies (e.g., clinical trials), reports with no usable information, and reports of non-brain imaging. We excluded people with records of stroke of any type, dementia, epilepsy, colorectal cancer, multiple sclerosis (MS), Parkinson's disease, transient ischaemic attack (TIA), subarachnoid haemorrhage, subdural haemorrhage and extradural haemorrhage, prior to the date of scan or 6 months after the date of scan (Supplementary Material). The process of selecting the analytic cohort is shown in Figure 1.

Study definitions

CCD and atrophy imaging phenotypes were white matter hypoattenuation (on CT) or hyperintensities (on MRI) (WMH), lacunes (small subcortical infarcts), cortical infarcts, and cerebral atrophy. Cerebral atrophy was included because it is a common co-finding of CCD and has potentially important disease associations. Rules-based NLP identified CCD or atrophy on first-available MRI or CT brain imaging report for each participant (https://www.ltg.ed.ac.uk/software/EdIE-R/). The tool has been validated against expert reads of Scottish brain imaging reports, and against an expert independent rater reading underlying images[6,9] Additionally, we created a variable summing the number of CCD and atrophy phenotypes present at the first brain image.

We defined outcomes as the first recorded date of: any stroke (and subtypes ischaemic stroke, intracerebral haemorrhage and uncertain stroke); all dementia (and subtypes Alzheimer's dementia, vascular dementia, and unspecified or rare dementia); Parkinson's disease; epilepsy; and colorectal cancer. Epilepsy and colorectal cancer was included as brain-based and non-brain negative control outcomes, respectively. Outcomes were defined with established ICD10 coding and British National Formulary (BNF) medication lists (Supplementary Material) taken from the Health Data Research UK Phenotype Library (https://phenotypes.healthdatagateway.org/) with SMR01, SMR04, SMR06, PIS and national death records.

Covariates included age (in years) at first-available imaging report date, split into 5-year bands, sex, area-level measures of deprivation at first brain imaging (Scottish Index for Multiple Deprivation (SIMD) decile (2012 classification),[11] Urban-Rural Classification (8-fold 2011 classification)[12] and Health Board of scan (a proxy for region). Ethnicity data was not available for the cohort.

We used all data sources to define co-morbidities prior to first imaging: myocardial infarction (MI), angina, peripheral vascular disease, renal impairment, heart failure, non-colorectal cancer and diabetes (Supplementary Material). We measured healthcare usage in the year prior to first scan with the number of hospital admissions, the number of unique medicines dispensed, the number of statin prescriptions dispensed, and the number of antihypertensive prescriptions dispensed.

Follow-up started from 6 months after the date of the first brain imaging report (index date), to allow time for diagnosis made with or because of the first brain imaging to appear in records. For each outcome separately, survival time was estimated to an event, or right-censored on death, migration out of Scotland, or study end (31/12/2020). We report 1-year, 5-year, and overall cumulative incidence with the Kaplan-Meier method for each outcome by scan phenotypes (WMH, lacunes, cortical infarcts, cerebral atrophy), and survival estimates relative to the wider Scottish population (the analytic cohort plus the unscanned population alive on 01/01/2010; N = 5,114,694). Survival analyses were conducted with Cox proportional hazards models and the proportional hazards assumption was tested statistically and visually with Schoenfeld residual plots. The variables 'scan modality', and the Scottish Health Board violated this assumption and were included as stratification variables. For each outcome (and its subtypes), and for each CCD and atrophy phenotype we estimated maximally adjusted hazard ratios (aHRs), adjusting for other phenotypes, age, sex, deprivation, rurality, co-morbidities and healthcare usage measured by prescriptions and admissions in the year prior to scan. We also examined adjusted HRs for number of CCD and atrophy phenotypes at scan (max 4). Full model estimates, including univariate HRs, are shown in Supplementary Material.

Sensitivity analyses for each outcome (and subtype) tested whether diagnosis preceded scan by excluding the first 1 and 5 years of follow-up. While the first 6 months of follow-up are excluded in the primary analyses, excluding larger time periods provides a stricter test of potential asymptomatic time, particularly for conditions with more insidious onset. Subgroup analyses tested aHR by age, sex, and brain imaging type.

Analyses were conducted within the NHS Scotland National Safe Haven, using R (v4.3.2)[13] and R Studio (v2023.12.0)[14] with the package 'survival'.[15] Missing data were not imputed.

The study was approved by the NRES Committee Northwest - Greater Manchester East ethics committee (15/NW/0719), and the Public Benefit and Privacy Panel for Health and Social Care (1516-0219) of NHS Scotland.

Role of the funding source

The funder of the study had no role in the design of the study, data collection, analysis, interpretation or writing of the report.

Results

Of 5·67 million people in Scotland between 2008 and 2018, 0·65 million (11%) had at least one brain imaging report, of whom 0·39 million (61%) had no recorded history of neurological disease. Of this 0.39 million, most (56%) were over 55 years old, 24% had a report of WMH, 8% of a lacune, 2% of a cortical infarct, 23% of cerebral atrophy and 63% had none of these scan phenotypes. There was a total of 2·8-million-person years of follow-up time, with

median follow-up time of people with CCD or atrophy of 5·8 years, and of scanned individuals without CCD or atrophy of 7·5 years (Supplementary Material).

Compared with people without, those with CCD or atrophy were older (mean 74·1 vs 48·1 years), had more admissions with MI (14·4% vs 4·2%), renal impairment (11·2% vs 2·4%), heart failure (6·5% vs 1·2%), cancer (14·2% vs 7·6%), and diabetes (10·5% vs 3·7%), had a greater number of hospitalisations (mean 1·4 vs 0·8) and prescribed medicines (3·2 vs 1·7) in the year prior to scan (Table 1). Compared with the Scottish population (with or without a brain imaging), people who had brain imaging were older, had more vascular and cancer comorbidities, and had fewer antihypertensive and statin dispensing events in the year prior to scan.

The five most frequent combinations of CCD and atrophy phenotypes were: WMH and cerebral atrophy; cerebral atrophy alone; WMH alone; WMH, cerebral atrophy and lacunes; and lacunes and WMH. The proportion of people who had an MRI rather than a CT was higher in people with no phenotype and for WMH only, although for all phenotype combinations, most brain imaging was performed with CT (Figure 2).

The incidence rate, absolute risk difference, and relative risk (RR) of outcomes between people with a brain image report and the wider Scottish population are shown in Supplementary Material. Incidence was proportionally higher in those with a brain image report than in the wider population for all outcomes except colorectal cancer, though RR were lower with increasing age (Supplementary Material). At one and five years, people with a CCD or atrophy phenotype were more likely than the wider Scottish population to experience stroke (1-year: RR 4·1, 95%CI: 3·9–4·4; 5-year: RR 33·2, 32·4–34·0), dementia (3·7, 3·6–3·8; 18·0, 17·8–18·3), Parkinson's disease (2·2, 2·0–2·4; 19·5, 18·7–20·3) and epilepsy (1·2, 1·2–1·3; 8·1, 8·0–8·2) but not colorectal cancer (0·7, 0·6–0·8; 5·4, 5·2–5·7). At one year, the cumulative incidence (accounting for the competing risk of death) of stroke and dementia were similar and low among those with WMH, lacunes, cortical infarcts, cerebral atrophy or no scan phenotype (around 1% and 5%; Table 2). At five years, cumulative stroke incidence was highest among those with cortical infarcts (9%), followed by lacunes (7%), WMH (6%) and cerebral atrophy (5%), while dementia incidence was highest among those with cerebral atrophy (22%), followed by WMH (20%), cortical infarcts (18%) and lacunes (18%).

Compared to those with no CCD or atrophy phenotype, those with each phenotype (accounting for the presence of other phenotypes) had a higher risk of any stroke over the 12-year follow-up period. The aHRs were, for cortical infarct 1·7 (1·6–1·8), lacunes 1·6 (1·5–1·6), WMH 1·4 (1·3–1·4), and cerebral atrophy 1·1 (95%CI: 1·0–1·1) (Figure 3). The aHR for ischaemic stroke was highest in people with cortical infarct (1·7, 1·6–1·9) and aHR for intracerebral haemorrhage was highest in people with cortical infarct (1·6, 1·5–1·8) and lacunes (1·6, 1·5–1·7). People with cerebral atrophy had an increased risk of unspecified stroke (1·2, 1·1–1·3), but not ischaemic stroke (1·0, 0·9–1·1) or haemorrhagic stroke (1·1, 0·9–1·1) (Figure 4). People with more CCD or atrophy phenotypes had a higher aHR for all stroke (1·3, 1·3–1·4), with the highest aHR (versus those without phenotypes) in those with all four phenotypes (2·9, 2·4–3·4) (Supplemental Material).

Compared with people with no evidence of CCD or atrophy, the aHR for dementia was higher in people with cerebral atrophy (aHR 1·7, 95%CI: 1·7–1·7), WMH (1·3, 1·3–1·3), and cortical infarct at scan (1·1, 1·0–1·1), although not lacunes (1·0, 0·9–1·0) (Figure 3). Of the CCD or atrophy phenotypes, only people with cerebral atrophy at scan had an increased risk of each dementia type: Alzheimer's disease (1·9, 1·8–2·1), vascular dementia (1·4, 1·3–1·4), and unspecified dementia (1·5, 1·4–1·6). Compared to those without CCD or atrophy

phenotypes, people with cortical infarct or lacunes had a higher risk of vascular dementia (cortical infarct: 1·7, 1·6–1·9; lacunes: 1·6, 1·5–1·7) but lower incidence of Alzheimer's disease (cortical infarct: 0·6, 0·5–0·8; lacunes: 0·7, 0·7–0·8). Compared to those without WMH, people with WMH had a higher risk of vascular dementia (2·0, 1·9–2·1) and unspecified dementia (1·3, 1·3–1·4), but not clearly Alzheimer's dementia (0·9, 0·9–1·0) (Figure 4). People with more CCD phenotypes had a higher dementia aHR (1·3, 1·3–1·3), with the highest aHR among those exhibiting all four phenotypes (2·5, 2·2–2·8).

Compared with people with no evidence of CCD or atrophy, the aHR of Parkinson's disease was lower in people with cortical infarct (aHR 0·7, 95%CI: 0·6–0·9) and higher in people with cerebral atrophy (1·4, 1·3–1·5) and WMH (1·1, 1·0–1·2), but not clearly higher in people with lacunes (1·1, 0·9–1·2) (Figure 3). There was a linear association between more scan phenotypes and Parkinson's disease risk (1·2, 1·1–1·2), but risk in those with all four phenotypes did not significantly differ to those with no phenotypes (1·0, 0·6–1·6). None of the CCD or atrophy phenotypes had a large aHR with risk of epilepsy (aHRs 0·9–1·0) or colorectal cancer (aHRs 0·9–1·1), and all had a p-value of >0.05.

Sensitivity analyses

After excluding the first year and five years of follow up, most estimates were of a similar magnitude and direction to the primary analyses. The exceptions were, after excluding five years of follow up, the aHR for cerebral atrophy with stroke risk was <1 and aHRs between lacunes and cortical infarcts with dementia were <1 (Supplementary Material), indicating that stroke and dementia are diagnosed earlier among these individuals.

Subgroup analyses

We tested subgroups by sex, age group, and scan modality. The aHR between stroke and WMH was higher in males than females ($p_{interaction} < 0·001$) and the aHRs between stroke and both cortical infarcts and atrophy were higher in females than males ($p_{interaction} <0·01$ and $p_{interaction} < 0·001$). The aHRs for WMH, lacunes, cortical infarcts and atrophy were stronger in younger than older people (all $p_{interaction} < 0·001$), and aHRs for lacunes and cerebral atrophy were stronger with MRI than CT imaging ($p_{interaction} < 0·001$). For dementia, aHRs for cerebral atrophy were stronger in males than females ($p_{interaction} < 0·05$), aHRs for WMH, cortical infarcts and cerebral atrophy were stronger in younger than older people (all $p_{interaction} < 0·001$), and aHRs for cerebral atrophy were stronger for MRI than CT ($p_{interaction} < 0·001$) (Supplementary Material).

Discussion

This cohort study of 395,273 people demonstrates that those with clinically reported CCD or cerebral atrophy have a higher incidence of future stroke and dementia diagnosis. Compared with people with normal scans, those with cortical infarct had the highest aHR of ischaemic stroke, and those with lacunes had the highest aHR of haemorrhagic stroke. Those with cerebral atrophy had the highest aHR of Alzheimer's disease, and those with WMH the highest aHR of vascular dementia. Compared with the general population (and those with normal scans), the absolute incidence of dementia and stroke was substantially elevated in people with these scan reports.

As expected, the prevalence of CCD and atrophy phenotypes in this clinical cohort was higher than in research cohorts, where MRI-discovered WMH was present in ~10% and silent brain infarcts in around ~2% of 50–69 year-olds (though prevalence increases with

age),[1] compared with 24% and 8% in our study. The aHRs of WMH and deep brain infarct with stroke and dementia in research cohort studies were also stronger (aHR>2) than the maximally adjusted estimates from the present study.[4] However, the aHRS reported here are similar to clinical studies for both stroke (aHRs 1·07–1·45)[16] and dementia (aHRs 1·28–1·93).[17] Lower aHRs in clinical than research cohort studies could be due to lack of precision in measurement of brain imaging phenotypes in clinical compared with research studies, or higher risk of death in people in clinical studies hence lower measured disease incidence. Overall, the observed aHRs between cortical infarcts and all stroke risk (1·7) and between cerebral atrophy and all dementia risk (1·7) are concordant with previous studies in non-stroke and post-stroke populations, consistent with the view that stroke involves more discrete vascular lesions while dementia involves more brain tissue loss.

Our study had several strengths. First, a validated rules-based NLP pipeline was used to detect markers of CCD and atrophy in clinical radiology reports. Text radiology reports are ideally suited for high-performance NLP, due to their replicable structure, limited phenotypes and short length. Previous studies have not measured multiple phenotypes, have used smaller cohorts (Ns of 261,960 and 241,050), and only examined older people.[16,17] Second, we took a whole clinical population approach, including both academic centres, where almost all research is currently done, and district hospitals, which do not usually contribute to research. Very few studies of imaging phenotypes have sample sizes of more than a thousand, and one of the largest cohorts, UK Biobank, is aiming for imaging one hundred thousand participants.[18] The large sample in the present study had sufficient statistical power to study subgroups and rare conditions such as Parkinson's disease. Third, linking participants to health systems data allowed ascertainment of multiple outcomes at low cost.[19–21] Fourth, we included an ageing-related non-brain condition (colorectal cancer) that was not associated with CCD or atrophy, increasing confidence that the findings were not entirely due to residual confounding.

Our study had several limitations. First, rules-based NLP models are not state-of-the-art in the field relative to large language models (LLM). However, rules-based models tend to perform well,[22] and the time taken for data governance approvals and limited compute resource meant that an LLM approach was not feasible when the project started. Second, our population was clinical, hence imaging was only performed in people at higher risk; this selection bias could limit generalisability to research cohorts or asymptomatic screening. Furthermore, imaging request information was too brief to meaningfully classify reports by indication, therefore stroke or dementia as a reason for imaging request could not be entirely excluded. However, we excluded people with a diagnosis of stroke or dementia prior to, or within 6 months of scanning and in sensitivity analysis found that extending this exclusion to one and five years made little qualitative difference to the pattern of aHRs. Third, we were only able to link to hospital, prescription, and death data, and did not have access to data from primary care, so some chronic conditions will have been under ascertained. This also meant that important confounders, such as obesity or smoking, were not available, and residual confounding could explain some of our findings. Fourth, there may have been diagnosis inclusion bias for dementia subtypes, because evidence of CCD could have increased the chance of a clinical diagnosis of vascular rather than other dementia subtypes. Lastly, our imaging phenotypes were relatively crude and did not include quantification or location. However, UK clinical brain imaging reports generally do not have sufficient detail to estimate the severity of WMH or atrophy, or precise location of deep or cortical infarcts. The appearance of old deep infarcts ('lacunes') is similar to perivascular spaces, recent small subcortical infarcts, small subcortical haemorrhage, or end-stage cavitation in a white matter hyperintensity, and is often difficult to distinguish in a clinical image. Ongoing work analysing the images in this dataset will provide quantitative phenotyping.

We have demonstrated that NLP can uncover previously difficult to discover phenotypes in electronic health records at a scale and speed not previously possible. However, care is needed as NLP can reflect systematic biases in the clinical text, for example incompleteness in reports in acute illness, and trends in language used in reports over time.[23–27] NLP phenotypes could be used to identify non-coded outcomes in large cohort studies that rely on the electronic health record for outcome ascertainment, or to identify populations eligible for clinical trials.[28] This may be true particularly for asymptomatic, incidentally discovered findings such as CCD.

This study emphasises that CCD or atrophy discovered during the process of routine healthcare are risk markers for stroke and dementia. In a clinical population, a report of atrophy indicates a modestly increased risk of subsequent dementia but does not make a clinical diagnosis, hence further assessment with cognitive testing may be needed. Apparently asymptomatic cortical infarcts or lacunes indicate a modestly increased risk of subsequent stroke, and in the absence of evidence to support specific interventions, careful application of primary vascular prevention – chiefly management of hypertension and high LDL cholesterol – should be the mainstay of treatment. Recent guidelines from the European Stroke Organisation (ESO) state that the best practice to prevent stroke or dementia in people with CCD remains uncertain.[29]

In conclusion, we implemented an NLP algorithm across a whole population and found that CCD and atrophy were associated with a higher risk of both dementia and stroke up to 12 years later, with important differences between CCD exposures. Clinical trials to reduce risk in people with CCD are warranted.

Contributors

WNW gained funding for the study. MHI conducted the statistical analysis. WNW, BA, CG, RT and ED implemented the natural language processing algorithm and extracted phenotype data. WNW, HW, MM, ED, HZ, LS and AH provided methodological and statistical advice. MHI and WNW drafted and redrafted the manuscript. All authors reviewed the manuscript and provided feedback.

Data Sharing

Data is restricted as per Public Health Scotland guidelines, as it relates to individual health records. Data can be made available within the National Safe Haven upon application and approval by Public Health Scotland and the Public Benefit and Privacy panel.

Funding

The project was funded by the Chief Scientist's Office (CSO-SCAF/17/01), the Medical Research Council (G0902303/1), the Alzheimer's Society (486). WW is supported by CSO and Health Data Research UK. MHI is supported by the Wellcome Trust (220857/Z/20/Z; 226770/Z/22/Z, 104036/Z/14/Z; 216767/Z/19/Z) and by a Research Data Scotland Accelerator Award (RAS-24-2). ELB is supported by MQ – Transforming Mental Health (MPSIP\30). BA is supported by the Turing Fellowship and Turing project (EP/N510129/1) from The Alan Turing Institute, by Legal and General PLC as part of the Advanced Care Research Centre, and by the National Institute for Health Research (NIHR202639). GM is supported by the Stroke Association Edith Murphy Foundation Senior Clinical Lectureship (SA L-SMP 18\1000). JMW is part funded by the UK DRI which is funded by UK MRC, Als Soc and ARUK; JMW also part funded by NIHR.

Declaration of interests

GM has received consultancy fees from Canon Medical Research, Europe, Ltd.

Acknowledgements

We would like to thank NHS Scotland patients for providing the data involved in this study. We would like to acknowledge the eDRIS team (Public Health Scotland) - particularly Rebecca Fairnie – for their support in obtaining approvals, the provisioning and linking of data and facilitating access to the National Safe Haven. We would also like to thank Arlene Casey for assistance setting up the NLP algorithm within the National Safe Haven.

References

1 Morris Z, Whiteley WN, Longstreth WT, *et al.* Incidental findings on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* 2009; **339**: b3016–b3016.

2 Vermeer SE, Longstreth WT, Koudstaal PJ. Silent brain infarcts: a systematic review. *Lancet Neurol* 2007; **6**: 611–9.

3 Smith EE, Saposnik G, Biessels GJ, *et al.* Prevention of Stroke in Patients With Silent Cerebrovascular Disease: A Scientific Statement for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke* 2017; **48**. DOI:10.1161/STR.0000000000000116.

4 Debette S, Schilling S, Duperron M-G, Larsson SC, Markus HS. Clinical Significance of Magnetic Resonance Imaging Markers of Vascular Brain Injury: A Systematic Review and Meta-analysis. *JAMA Neurol* 2019; **76**: 81.

5 Fry A, Littlejohns TJ, Sudlow C, *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017; **186**: 1026–34.

6 Alex B, Grover C, Tobin R, Sudlow C, Mair G, Whiteley W. Text mining brain imaging reports. *J Biomed Semant* 2019; **10**: 23.

7 Davidson EM, Poon MTC, Casey A, *et al.* The reporting quality of natural language processing studies: systematic review of studies of radiology reports. *BMC Med Imaging* 2021; **21**: 142.

8 Wheater E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. *BMC Med Inform Decis Mak* 2019; **19**: 184.

9 Casey A, Davidson E, Grover C, *et al.* Understanding the performance and reliability of NLP tools: a comparison of four NLP tools predicting stroke phenotypes in radiology reports. *Front Digit Health* 2023; **5**: 1184919.

10 Nind T, Sutherland J, McAllister G, *et al.* An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population. *GigaScience* 2020; **9**: giaa095.

11 Scottish Government. Scottish Index of Multiple Deprivation 2020: Introduction. 2020 https://www.gov.scot/publications/scottish-index-multiple-deprivation-2020/.

12 Scottish Government urban rural classification, 2016. Edinburgh: Scottish Government, 2018.

13 R Core Team. R: A Language and environment for statistical computing. 2023. https://www.r-project.org/.

14 RStudio Team. RStudio: Integrated development for R. 2023. http://www.posit.co/.

15 Therneau TM, Lumley T, Atkinson E, Crowson C. Survival Analysis. 2023. https://github.com/therneau/survival.

16 Wang AY, Leung LY, Puttock EJ, *et al.* Stratifying Future Stroke Risk with Incidentally-Discovered White Matter Disease Severity and Covert Brain Infarct Site. *Cerebrovasc Dis* 2023; **53**: 509–14.

17 Kent DM, Leung LY, Zhou Y, *et al.* Association of Incidentally Discovered Covert Cerebrovascular Disease Identified Using Natural Language Processing and Future Dementia. *J Am Heart Assoc* 2023; **12**: e027672.

18 Littlejohns TJ, Holliday J, Gibson LM, *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat Commun* 2020; **11**: 2624.

19 Wilkinson T, Ly A, Schnier C, *et al.* Identifying dementia cases with routinely collected health data: A systematic review. *Alzheimers Dement* 2018; **14**: 1038–51.

20 Wilkinson T, Schnier C, Bush K, *et al.* Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. *Eur J Epidemiol* 2019; **34**: 557–65.

21 Rannikmäe K, Ngoh K, Bush K, *et al.* Accuracy of identifying incident stroke cases from linked health care data in UK Biobank. *Neurology* 2020; **95**. DOI:10.1212/WNL.0000000000009924.

22 Sykes D, Grivas A, Grover C, *et al.* Comparison of rule-based and neural network models for negation detection in radiology reports. *Nat Lang Eng* 2021; **27**: 203–24.

23 Casey A, Davidson E, Poon M, *et al.* A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak* 2021; **21**: 179.

24 Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. *Ann Surg* 2018; **268**: 70–6.

25 Yim W, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. *JAMA Oncol* 2016; **2**: 797.

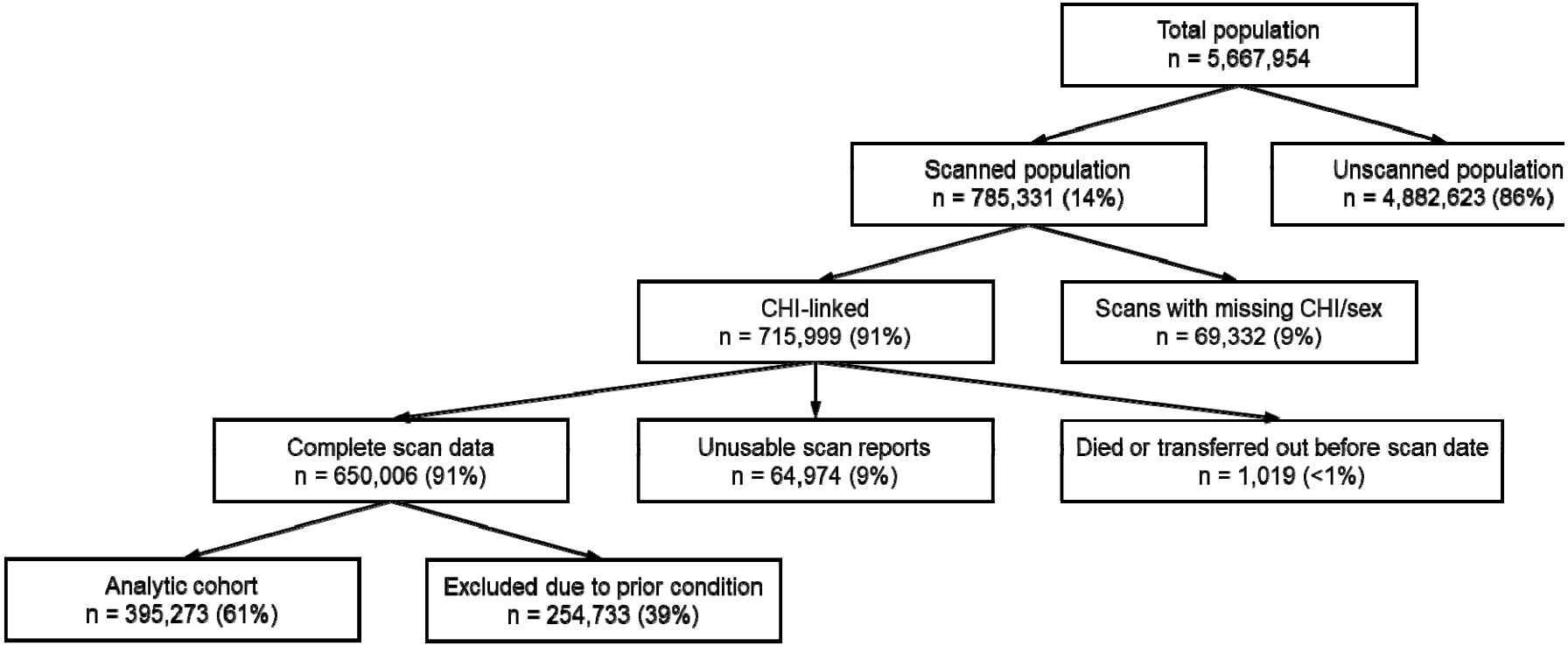
26 Baclic O, Tunis M, Young K, Doan C, Swerdfeger H. Challenges and opportunities for public health made possible by advances in natural language processing. *Can Commun Dis Rep* 2020; : 161–8.

27 Lan Z, Turchin A. Impact of possible errors in natural language processing-derived data on downstream epidemiologic analysis. *JAMIA Open* 2023; **6**: ooad111.

28 Kantor K, Morzy M. Machine learning and natural language processing in clinical trial eligibility criteria parsing: a scoping review. *Drug Discov Today* 2024; **29**: 104139.

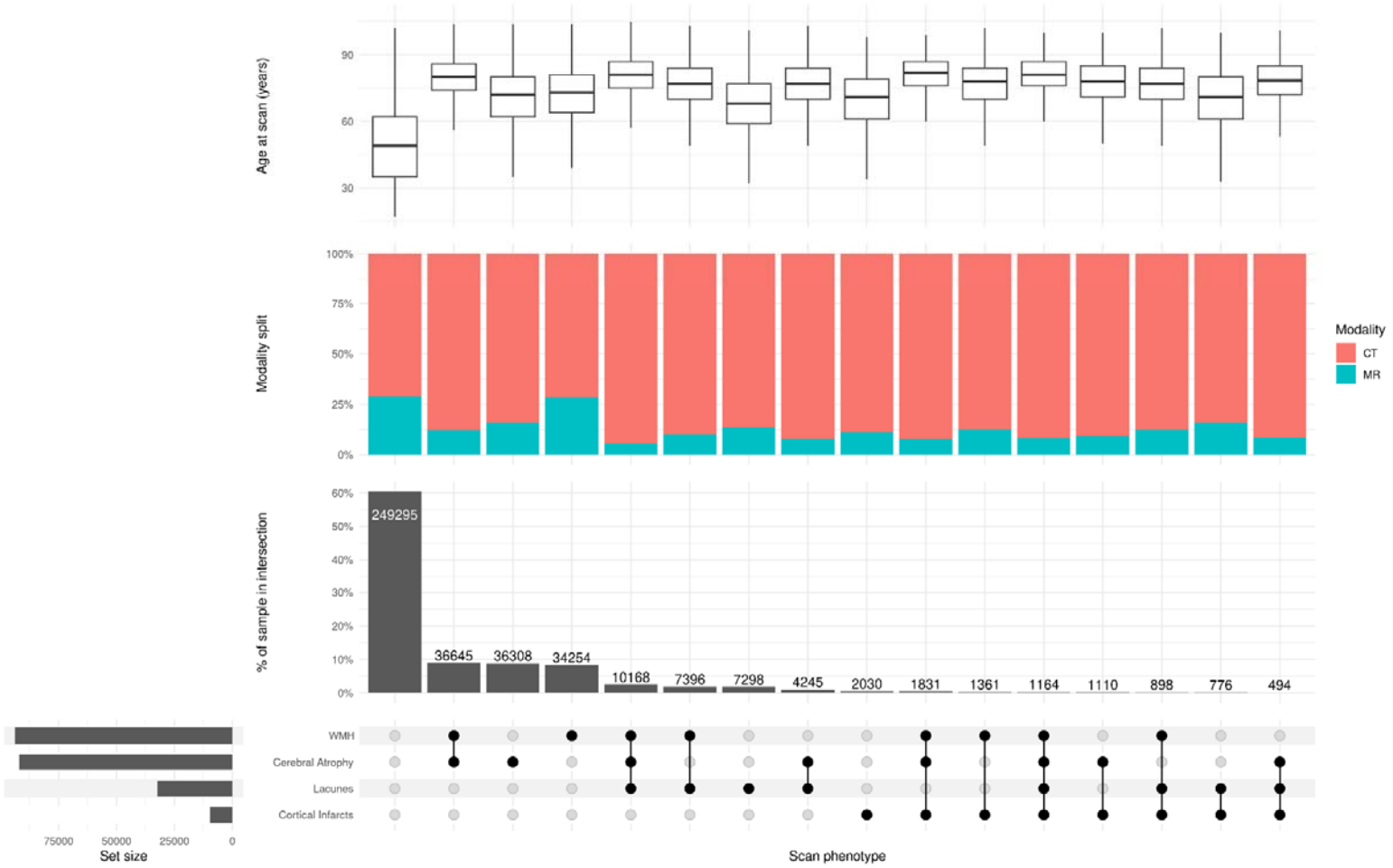
29 Wardlaw JM, Debette S, Jokinen H, *et al.* ESO Guideline on covert cerebral small vessel disease. *Eur Stroke J* 2021; **6**: CXI–CLXII.

Figure 1. Flow diagram showing cohort selection from the whole population of Scotland who interacted with the Scottish NHS (N = 5,667,954) to the analytic cohort (N = 395,273). Percentages indicate the proportion of the node above.



CHI = Community Health Index

Figure 2. Co-occurrence of covert cerebrovascular disease phenotypes, with summary of imaging modality and age at imaging per intersection. Box and whisker plots represent range, interquartile range and median.



WMH = white matter hypoattenuation/hyperintensities

Figure 3. Forest plot showing the event rate and adjusted hazard ratio (aHR) and 95% confidence intervals for each exposure (white matter hyper (MRI) or hypointensity (CT) (WMH), lacunes, cortical infarcts and cerebral atrophy) and their association with stroke risk, dementia risk, Parkinson's risk, epilepsy risk and colorectal cancer risk across follow-up.

Outcome	Events (n)	aHR	95% CI
Stroke	**11789**		
WMH	5925	1.36	1.31—1.42
Lacunes	2658	1.57	1.50—1.64
Cortical Infarcts	949	1.71	1.60—1.83
Cerebral Atrophy	5280	1.07	1.03—1.12
Dementia	**32614**		
WMH	18862	1.29	1.26—1.32
Lacunes	6129	1.02	0.99—1.05
Cortical Infarcts	1845	1.10	1.04—1.15
Cerebral Atrophy	19949	1.70	1.66—1.74
Parkinsons	**3558**		
WMH	1572	1.08	1.00—1.16
Lacunes	560	1.07	0.98—1.18
Cortical Infarcts	118	0.71	0.59—0.85
Cerebral Atrophy	1782	1.43	1.33—1.54
Epilepsy	**39751**		
WMH	7285	1.00	0.97—1.03
Lacunes	2527	0.96	0.92—1.00
Cortical Infarcts	764	0.98	0.91—1.06
Cerebral Atrophy	7295	1.01	0.98—1.04
Colorectal cancer	**2304**		
WMH	876	0.94	0.86—1.03
Lacunes	322	1.05	0.93—1.19
Cortical Infarcts	100	1.07	0.87—1.31
Cerebral Atrophy	897	1.02	0.93—1.12

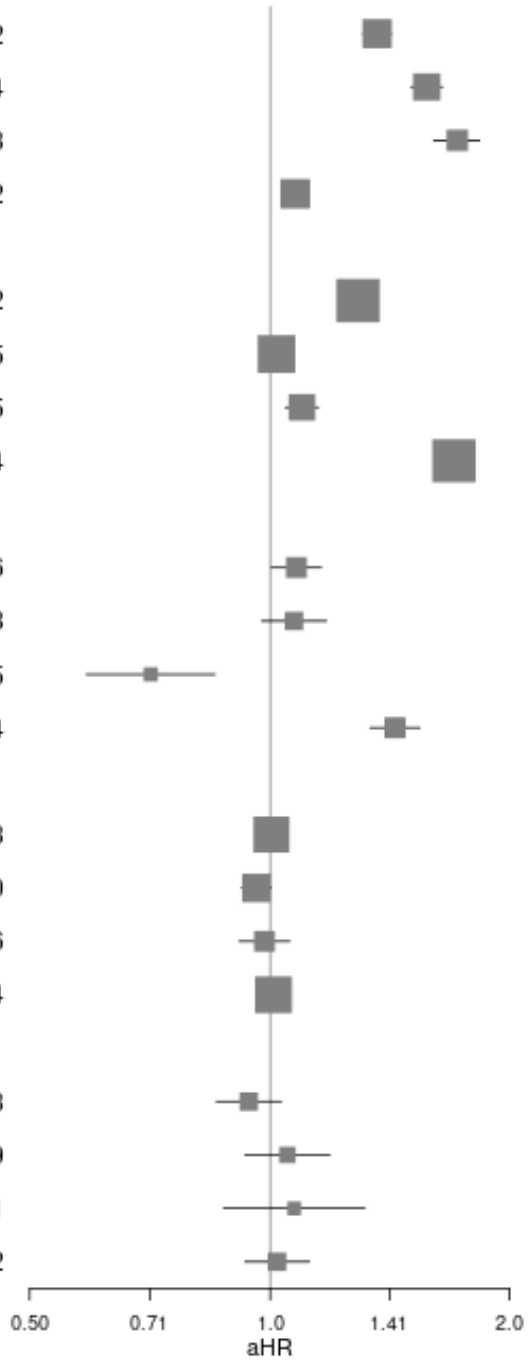
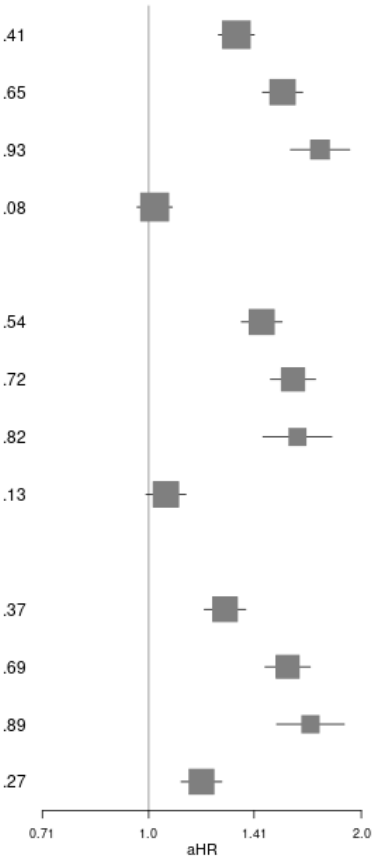


Figure 4. Forest plot showing the event rate and adjusted hazard ratio (aHR) and 95% confidence intervals for each exposure (white matter hyper (MRI) or hypointensity (CT) (WMH), lacunes, cortical infarcts, and cerebral atrophy) and their association with risk of a) stroke subtypes and b) dementia subtypes over follow-up.

Outcome	Events (n)	aHR	95% CI
Ischaemic Stroke	**5878**		
WMH	2805	1.3	1.26—1.41
Lacunes	1277	1.6	1.45—1.65
Cortical Infarcts	470	1.8	1.59—1.93
Cerebral Atrophy	2465	1.0	0.96—1.08
Haemorrhagic Stroke	**4550**		
WMH	2304	1.4	1.35—1.54
Lacunes	1023	1.6	1.49—1.72
Cortical Infarcts	346	1.6	1.45—1.82
Cerebral Atrophy	1992	1.1	0.99—1.13
Unspecified Stroke	**4271**		
WMH	2248	1.3	1.20—1.37
Lacunes	1015	1.6	1.46—1.69
Cortical Infarcts	361	1.7	1.52—1.89
Cerebral Atrophy	2121	1.2	1.11—1.27



Outcome	Events (n)	aHR	95% CI
Alzheimer's Dementia	**3075**		
WMH	1650	0.94	0.88—1.02
Lacunes	426	0.73	0.65—0.81
Cortical Infarcts	99	0.65	0.53—0.79
Cerebral Atrophy	2017	1.90	1.76—2.06
Vascular Dementia	**8578**		
WMH	5888	1.97	1.88—2.07
Lacunes	2489	1.60	1.53—1.68
Cortical Infarcts	831	1.73	1.61—1.86
Cerebral Atrophy	5247	1.37	1.31—1.43
Unspecified Dementia	**10363**		
WMH	6268	1.31	1.25—1.36
Lacunes	2028	1.02	0.97—1.07
Cortical Infarcts	644	1.16	1.07—1.25
Cerebral Atrophy	6390	1.49	1.43—1.56

Table 1. Demographic characteristics of those with and without clinically reported covert cerebrovascular (CCD) or cerebral atrophy phenotype and of the whole Scottish population (analytic cohort plus unscanned individuals alive at 01/01/2010)**.**

		Analytic cohort			Scottish population
		Covert cerebrovascular disease or cerebral atrophy n= 145,978	No covert cerebrovascular disease or cerebral atrophy n= 249,295	All N = 395,273	N = 5,114,694
Sex	Female	77,243 (52·9%)	135,052 (54·2%)	212,295 (53·7%)	2,690,704 (52·6%)
Age (years)	Mean (SD)	74·1 (12·5)	48·9 (17·5)	58·2 (19·9)	37·6 (25·1)
Age group	<=25	338 (0·2%)	26,740 (10·7%)	27,078 (6·9%)	616,056 (12·0%)
	26-35	927 (0·6%)	38,777 (15·6%)	39,704 (10·0%)	621,523 (12·2%)
	36-45	2773 (1·9%)	42,379 (17·0%)	45,143 (11·4%)	689,434 (13·5%)
	46-55	8730 (6·0%)	51,049 (20·5%)	59,779 (15·1%)	709,087 (13·9%)
	56-65	19,013 (13·0%)	42,695 (17·1%)	61,708 (15·6%)	597,648 (11·7%)
	66-75	38,368 (26·3%)	29,808 (12·0%)	68,176 (17·3%)	412,577 (8·1%)
	76-85	50,611 (34·7%)	14,175 (5·7%)	64,786 (16·4%)	244,363 (4·8%)
	86-95	23,711 (16·3%)	3,496 (1·4%)	27,207 (6·9%)	89,062 (1·7%)
	>=96	1,516 (1·0%)	176 (<0·1%)	1,692 (0·4%)	8,295 (12·2%)
Scan modality	CT	122,279 (83·8%)	177,487 (71·2%)	299,766 (75·8%)	-
	MR	23,699 (16·2%)	71,808 (28·8%)	95,507 (24·2%)	-
SIMD* quintile	1 Most deprived	34,107 (23·4%)	61,142 (24·6%)	95,249 (24·1%)	-
	2	30,670 (21·0%)	51,184 (20·5%)	81,854 (20·7%)	-
	3	26,893 (18·4%)	46,385 (18·6%)	73,278 (18·6%)	-
	4	25,848 (17·7%)	43,369 (17·4%)	69,217 (17·5%)	-
	5 Least deprived	25,798 (17·7%)	39,927 (16·0%)	65,725 (16·6%)	-
	Missing	2,662 (1·8%)	7,288 (2·9%)	9,950 (2·5%)	-
Urban-Rural**	1 Large Urban	60,383 (41·4%)	100,648 (40·4%)	161,031 (40·7%)	-
	2	43,538 (29·8%)	73,223 (29·4%)	116,761 (29·5%)	-
	3	12,064 (8·3%)	19,620 (7·9%)	31,684 (8·0%)	-
	4	3,477 (2·4%)	5,153 (2·0%)	8,630 (2·2%)	-
	5	1,243 (0·8%)	2,312 (0·9%)	3,555 (0·9%)	-
	6	14,442 (9·9%)	25,118 (10·1%)	39,560 (10·0%)	-
	7	4,099 (2·8%)	6,736 (2·7%)	10,835 (2·8%)	-
	8 Very Remote Rural	3,225 (2·2%)	5,677 (2·3%)	8,902 (2·3%)	-
	Missing	3,507 (2·4%)	10,808 (4·3%)	14,315 (3·6%)	-
Comorbidities	MI, angina, or peripheral vascular disease	21,024 (14·4%)	10,446 (4·2%)	31,470 (8·0%)	90,927 (1·8%)
	Renal impairment	16,361 (11·2%)	5,909 (2·4%)	22,270 (5·6%)	44,017 (0·9%)
	Heart failure	9,529 (6·5%)	3,019 (1·2%)	12,548 (3·2%)	31,668 (0·6%)
	Any cancer	20,795 (14·2%)	18,985 (7·6%)	39,780 (10·1%)	102,815 (2·0%)
	Diabetes	15,385 (10·5%)	9,170 (3·7%)	24,555 (6·2%)	621,523 (12·2%)
Care in year before scan	Hospital admissions (mean, SD)	1·4 (2·9)	0·8 (2·3)	1·0 (2·5)	0·9 (2·1)
	Unique medicines (mean, SD)	3·2 (2·7)	1·7 (2·2)	2·2 (2·5)	2·8 (2·5)
	Antihypertensive dispensing events (mean, SD)	2·4 (3·7)	0·8 (2·4)	1·4 (3·1)	3·7 (6·9)
	Statin dispensing events (mean, SD)	1·0 (1·7)	0·4 (1·1)	0·6 (1·4)	1·5 (3·1)

*2012 **2011-2012

Table 2. 1-year and 5-year cumulative incidence (%) for each scan phenotype and outcome, accounting for competing risk of death.

	WMH		Lacunes		Cortical infarcts		Cerebral atrophy		None		Analytic cohort		Scottish population	
Outcome	1-year	5-year	1-year	5-year	1-year	5-year	1-year	5-year	1-year	5-year	1-year	5-year	1-year	5-year
Stroke	0·9 [0·9-1·0]	5·6 [5·5-5·8]	1·2 [1·1-1·3]	7·2 [6·9-7·5]	1·6 [1·3-1·8]	8·6 [8·3-9·4]	0·8 [0·8-0·9]	5·1 [5·0-5·2]	0·1 [0·1-0·2]	1·0 [1·0-1·1]	0·4 [0·4-0·4]	2·5 [2·5-2·6]	0·1 [0·1-0·1]	0·4 [0·4-0·4]
Dementia	5·1 [5·0-5·2]	20·0 [19·0-20·0]	4·4 [4·2-4·6]	18·0 [18·0-19·0]	4·6 [4·2-5·0]	18·0 [18·0-19·0]	6·0 [5·9-6·2]	22·0 [21·0-22·0]	0·5 [0·4-0·5]	2·0 [2·0-2·1]	2·1 [2·0-2·1]	8·4 [8·3-8·5]	0·6 [0·6-0·6]	1·5 [1·5-1·5]
Parkinson's	0·2 [0·2-0·3]	1·6 [1·5-1·7]	0·3 [0·2-0·3]	1·6 [1·5-1·7]	0·1 [0·1-0·2]	1·1 [0·9-1·3]	0·3 [0·2-0·3]	1·8 [1·8-1·9]	<0·1 [<0·1-<0·1]	0·3 [0·3-0·4]	0·1 [0·1-0·1]	0·9 [0·8-0·9]	0·1 [<0·1-0·1]	0·2 [0·2-0·2]
Epilepsy	1·6 [1·5-1·7]	7·8 [7·6-7·9]	1·6 [1·4-1·7]	7·7 [7·4-8·0]	1·5 [1·3-1·8]	7·6 [7·1-8·2]	1·6 [1·5-1·6]	7·9 [7·7-8·1]	2·2 [2·2-2·3]	11·0 [11·0-1·4]	2·0 [2·0-2·1]	10·9 [10·8-11·0]	1·7 [1·7-1·7]	4·7 [4·7-4·7]
Colorectal Cancer	0·1 [0·1-0·2]	0·9 [0·8-0·9]	0·1 [0·1-0·2]	0·9 [0·8-1·0]	0·2 [0·1-0·3]	1·0 [0·8-1·2]	0·2 [0·1-0·2]	0·9 [0·8-1·0]	<0·1 [<0·1-0·1]	0·3 [0·3-0·3]	0·1 [0·1-0·1]	0·5 [0·5-0·6]	0·1 [0·1-0·1]	0·4 [0·4-0·4]

WMH = white matter hypoattenuation/hyperintensities