BMJ Open

Cohort profile: AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353,157 patients in London, United Kingdom

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-058552
Article Type:	Cohort profile
Date Submitted by the Author:	20-Oct-2021
Complete List of Authors:	Wagner, Siegfried; University College London Institute of Ophthalmology, Hughes, Fintan; Duke University Hospital Pontikos, Nikolas; University College London; Moorfields Eye Hospital NHS Foundation Trust Struyven, Robbert; University College London; Moorfields Eye Hospital NHS Foundation Trust Alexander, Daniel; University College London, Department of Computer Science Hindley, Jack; University College London, Department of Information Governance Keane, Pearse; University College London; Moorfields Eye Hospital NHS Foundation Trust Topol, Eric; Scripps Research Institute Cortina Borja, Mario; UCL, Institute of Child Health Liu, Xiaoxuan; University Hospitals Birmingham NHS Foundation Trust; University of Birmingham Montgomery, Hugh; University College London, Centre for Human Health and Performance Petersen, Steffen; Queen Mary University of London, William Harvey Research Institute; Barts Health NHS Trust, Balaskas, Konstantinos; Moorfields Eye Hospital NHS Foundation Trust; Moorfields Eye Hospital NHS Foundation Trust Petzold, Axel; UCL; Moorfields Eye Hospital NHS Foundation Trust Rahi, Jugnoo; UCL, Great Ormond Street Institute of Child Health; Great Ormond Street Hospital for Children NHS Foundation Trust Denniston, Alastair; Queen Elizabeth Hospital Birmingham, UK; University of Birmingham
Keywords:	OPHTHALMOLOGY, Medical ophthalmology < OPHTHALMOLOGY, Medical retina < OPHTHALMOLOGY, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Cohort profile: AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353,157 patients in London, United Kingdom

SK Wagner[1, 2], F Hughes[3], M Cortina-Borja[4], N Pontikos[1, 2], RR Struyven[1, 2], X Liu[5, 6], H Montgomery[7], DC Alexander[8], E Topol[9], S Petersen[10, 11], K Balaskas[1, 2], J Hindley[12], A Petzold[1, 2, 13, 14], J Rahi[1, 4, 15, 16], AK Denniston[5, 6], PA Keane[1, 2, *]

*Corresponding author: Pearse A. Keane, MD FRCOphth, NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, 162 City Road, London, United Kingdom. Tel: +44 207 253 3411 Email: pearse.keane1@nhs.net

Word count (excluding title page, abstract, references, figures and tables): 4915 words

[1] Institute of Ophthalmology, University College London, London, UK,

[2] Moorfields Eye Hospital NHS Foundation Trust, London, UK,

[3] Duke University Hospital, Durham, North Carolina, USA,

[4] Great Ormond Street Institute of Child Health, University College London, London, UK,

[5] University of Birmingham, Birmingham, UK,

[6] University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK,

[7] Centre for Human Health and Performance, University College London, London, UK

[8] Centre for Medical Image Computing, Department of Computer Science, University College London, UK

[9] Department of Molecular Medicine, Scripps Research, La Jolla, CA, USA.

[10] William Harvey Research Institute, Queen Mary University of London, London, UK.

[11] St Bartholomew's Hospital, Barts Health National Health Service (NHS) Trust, London, UK..

[12] Department of Information Governance, School of Life and Medical Sciences, University College London, London, UK

[13] The National Hospital for Neurology and Neurosurgery (UCLH), Queen Square Institute of Neurology, University College London, London, UK.

[14] Neuro-ophthalmology Expert Centre, Amsterdam UMC, Departments of Neurology and Ophthalmology, Netherlands.

[15]Great Ormond Street Hospital NHS Foundation Trust, London, United Kingdom.

[16]Ulverscrosft Vision Research Group, University College London, London, UK

Abstract

Purpose: Retinal signatures of systemic disease ('oculomics') are increasingly being revealed through a combination of high-resolution ophthalmic imaging and sophisticated modelling strategies. Progress is currently limited not mainly by technical issues, but by the lack of large labelled datasets, a sine qua non for deep learning. Such data are derived from prospective epidemiological studies, in which retinal imaging is typically unimodal, cross-sectional, of modest number and relate to cohorts, which are not enriched with subpopulations of interest, such as those with systemic disease. We thus linked longitudinal multimodal retinal imaging from routinely-collected National Health Service data with systemic disease data from hospital admissions using a privacy-by-design third-party linkage approach.

Participants: Between January 1st 2008 and April 1st 2018, 353,157 participants aged 40 years or older, who attended Moorfields Eye Hospital NHS Foundation Trust, a tertiary ophthalmic institution incorporating a principal central site, four district hubs and five satellite clinics in and around London, United Kingdom serving a catchment population of approximately six million people.

Findings to date: Among the 353,157 individuals, 186,651 had a total of 1,337,711 Hospital Episode Statistics admitted patient care episodes. Systemic diagnoses recorded at these episodes include 12,022 patients with myocardial infarction, 11,735 with all-cause stroke and 13,363 with all-cause dementia. A total of 6,261,931 retinal images of seven different modalities and across three manufacturers were acquired from 154,830 patients. The majority of retinal images were retinal photographs (n=1,874,175) followed by optical coherence tomography (n=1,567,358).

Future plans: AlzEye combines the world's largest single institution retinal imaging database with nationally collected systemic data to create an exceptional large-scale, enriched cohort that reflects the diversity of the population served. First analyses will address cardiovascular diseases and dementia, with a view to identifying hidden retinal signatures that may lead to earlier detection and risk management of these life-threatening conditions.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- Retinal signatures of systemic diseases, particularly cardiovascular and neurodegenerative diseases, have highlighted the eye as a potential window to risk-stratification.
- AlzEye is a large curated dataset linking more than six million routinely collected retinal images with National Health Service hospital admissions data over a ten-year period in 353,157 patients.
- The cohort is ethnically and socioeconomically diverse and consists of an enriched subpopulation of 12,022 individuals with myocardial infarction and 13,363 individuals with all-cause dementia.
- AlzEye provides a powerful foundation for the development and validation of both traditional and deep learning-based static and dynamic risk prediction models of systemic disease from retinal morphology.

INTRODUCTION

Scientific discovery has increasingly been driven by the availability of large, diverse, high-dimensional datasets providing deeply phenotyping variables in health and disease[1–3]. Advances in healthcare informatics, hardware and statistical techniques have uncovered relationships previously unachievable through traditional methods of study design. One exemplar is genomics in which the rich and voluminous data afforded by modern sequencing technology has highlighted the contribution of the genome to the trajectory of specific diseases and therapeutic response[4,5]. Similarly, the dynamic field of radiomics seeks to explore meaningful relationships between quantitative data extracted from medical imaging and disease at substantial scale[6].

A crucial substrate for health data research has been the establishment of and accessibility to large prospective epidemiological studies, such as the United Kingdom Biobank (UKBB), the Rotterdam study, and the European Prospective Investigation into Cancer and Nutrition (EPIC) study[7–9]. While such studies undoubtedly represent an exceptionally powerful enabler for discovery science, as evidenced by the high number of publications drawing on them[10], they are potentially limited for investigations of specific subpopulations of interest (e.g. those with rare disease or specific sociodemographic groups) and, where they draw on volunteer participants, also prone to selection bias (e.g. over-representation of more healthy subjects). For example, participants in UKBB are less likely to be obese, smoke, or drink alcohol and accordingly, mortality rates for participants aged 70-74 in UKBB are 46.2% and 55.5% lower for men and women respectively compared to general UK population[11].

Healthcare in England is such that when a patient has a medical event requiring admission, in almost all cases they are admitted under the provisions of the National Health Service (NHS).

Following discharge, routinely collected healthcare administrative data during a patient's admission are subsequently translated into corresponding International Classification of Diseases (ICD) codes by clinical coders within the respective institution and submitted to the Secondary Uses Service. This process is overseen by NHS Digital, who provide a unified record-level national repository of Hospital Episode Statistics (HES) data relating admitted patient care (APC) for almost all of the English population. While the original purpose of HES was the monitoring of service activity and negotiation of financial reimbursement, it is increasingly used for epidemiological research[12]. HES data are amenable to research as a sole resource. However, using deterministic linkage where individual identifiers are matched in a rules-based approach in contrast to probabilistic linkage[13], HES can enrich other datasets as in the case of UKBB[14], the Avon Longitudinal Study of Parents and Children[15] or the European Prospective Investigation into Cancer in Norfolk[16].

While these aforementioned studies demonstrate the value of enriching structured datasets through linkage with HES, this has not yet been done at scale for routinely collected data of high-dimensionality, such as imaging. Thus, we established AlzEye, a large linked dataset which combines routinely collected retinal images and relevant ophthalmic data from an unselected population attending Moorfields Eye Hospital NHS Foundation Trust (MEH), and nationally collected systemic healthcare outcome data provided through the HES APC data record of NHS Digital. MEH is a tertiary ophthalmic institution incorporating a principal central site, four district hubs and five satellite clinics in London, United Kingdom (UK), providing care to an ethnically and socioeconomically diverse population of six million people (9% of the UK population)[17]. The primary aim of AlzEye is to characterise the association between imaging-derived retinal biomarkers and chronic complex disorders of ageing, particularly dementia and cardiovascular diseases. In addition to describing the characteristics of the AlzEye cohort and dataset, in this paper, we describe the key governance, technical and ethical factors that need

to be addressed to enable large institution-led individual-level linkage of routinely collected multi-dimensional data. We also share an approach that has enabled us to create an exceptional trans-disciplinary resource that can support sophisticated modelling approaches, such as deep learning, so as to explore the retinal signatures of systemic disease.

COHORT DESCRIPTION

Study population

The AlzEye project is a retrospective cohort study of patients aged 40 years and over who have attended MEH between January 1st 2008 and April 1st 2018. Patients were included if they had attended the glaucoma, retina, neuro-ophthalmology or emergency ophthalmic services and had valid NHS numbers. Those with invalid NHS numbers, dates of birth or who had previously opted out of their health data being used for purposes of research (described in the NHS as a 'Type 2 opt-out') were excluded. Ethnicity group was self-reported by the patient as i) Asian or Asian British, ii) Black or Black British, iii) Mixed, iv) Other Ethnic Group, v) White, or vi) Unknown. Socio-economic status was categorised using the Index of multiple deprivation (IMD) decile, which was estimated by permuting the IMD 2015 rank from the patient's postcode through Lower Super Output Areas followed by aggregation into deciles[18]. Mortality data were derived from the MEH database, which is updated on a two-weekly basis using reports extracted from the NHS National Spine, and is completed on an individual basis by the MEH Data Quality team to ensure accuracy. Data are completed on any patients who have ever attended MEH. Mortality data up to the end of the study period, 1st April 2018, were included.

Approvals and process

The following key steps in the governance processes were required to establish AlzEye, and provide the necessary ongoing assurance within the research ethics framework of the NHS and the legal framework of the UK. In order to support other researchers who wish to establish similar linked cohorts, we provide an explanation of each stage which outlines the principle which that stage addresses, the UK framework that meets that principle, and finally any study-specific considerations that we undertook to not only meet but exceed those requirements (Figure 1).

- Funding

It was necessary to secure funding to deliver the study, and to provide assurance to the sponsor and others that the study would be completed and that the integrity of the study would not be compromised by inadequate resources. In AlzEye, the study was funded through a small grant awarded by Fight for Sight and Alzheimer's Research UK in October 2017 covering the costs of data storage and linkage fees. The funders had no role in the conception, design or analysis of the study.

- Sponsorship

It was necessary to secure a sponsor for the study that would take on 'overall responsibility for proportionate, effective arrangements being in place to set up, run and report on a research project'. In AlzEye, we sought sponsorship from the relevant NHS Trust (MEH) at which all patients had been seen. Sponsorship confirmation was only sought following internal consultation between data protection, information governance, information security and information technology (IT) teams at both MEH and UCL. MEH acted as the Sponsor, with UCL acting as the trusted third party linking retinal images and HES data and providing

computational facilities for data analysis. The study and data governance were approved on 24/05/2018 (internal reference: KEAP1004).

- Health Research Authority (HRA) Approval

For most research studies in England and Wales, including those limited to working with data for specific projects, Health Research Authority (HRA) approval is required. HRA approvals involve the assessment of governance and legal compliance of a research study with an independent ethics review and opinion from the Research Ethics Committee (REC). Depending on other study characteristics (e.g. the use of ionising radiation, gene therapy), additional applications may be required to inform the HRA approval. In England, limited access to confidential patient information without consent may be granted under the provisions of Section 251 of the National Health Service Act 2006. This permits temporary lifting of the common law duty of confidentiality around confidential patient information 'in the public interest' or 'in the interests of improving patient care'[19]. Obtaining section 251 support requires application to the Confidential Advisory Group (CAG), an independent body providing expert advice to the HRA for research applications and NHS Digital for data dissemination. Applications were accordingly made to the Research Ethics Committee (18/LO/1163, approved 01/08/2018) and the CAG for Section 251 support (18/CAG/0111, approved 13/09/2018). The National Health Service Health Research Authority gave final approval on 13/09/2018. Approvals thus far granted the legal basis for submitting an application to the Data Access Request Service (DARS) of NHS Digital[20].

- NHS Digital and the Data Access Request Service (DARS)

Among its services as the national information and technology partner for the NHS, NHS Digital oversees the Data Access Request Service (DARS), which administers and provides, upon application, multiple England-wide datasets from disease-specific audits (e.g. National Diabetes

Audit Core) to general admissions in secondary care (e.g. Hospital Episode Statistics). Applications to DARS require that the organisation have, at a minimum, the following:

i) Data Sharing Framework Contract for Data Controllers

ii) Compliance with minimum-security standards for Data Processors and Data Storage locations

iii) Adequate information security certification (e.g. ISO27001)

iv) A legal basis for data access (e.g. Section 251)

Applications are then reviewed with an assigned case officer, who will liaise with the applicant on project-specific items. For AlzEye, dialogue between the applicant and NHS Digital data production team revolved around confirmation of data fields and datasets (HES) and the pseudonymisation embedded within the linkage strategy.

- Independent Group Advising on the Release of Data (IGARD)

Following internal NHS Digital review and prior to data release, DARS applications are scrutinised by the Independent Group Advising on the Release of Data (IGARD) in line with Section 263(2) of the Health and Social Care Act 2012, the Code of Practice on confidential information. IGARD is an independent panel with a broad range of expertise, from legal to information governance to epidemiology. Support for AlzEye was given by IGARD in January and August 2019 citing that 'aspects of the application could be used as an exemplar by NHS Digital to help other researchers with their applications to the Data Access Request Service (DARS)'[21].

- Data sharing agreements (DSA)

Prior to data receipt, data sharing agreements (DSA) must be signed between NHS Digital and the Data Controller and are overseen by their respective legal departments. AlzEye required an

additional DSA between MEH and UCL for the transfer of ophthalmic imaging and clinical data between institutions outlining the purpose and legal basis for sharing.

- Data processing and transfer

The dataset was finalised upon completion of engineering work parsing manufacturer-specific file formats to non-proprietary data structures amenable to image analysis with appropriate deidentification. A secure cloud-based informatics pipeline was used for transfer of images to UCL from MEH, the establishment of which was delayed by the COVID-19 pandemic. Imaging data was stored (with backup) across four dedicated discs on a network-attached storage device within the UCL Institute of Ophthalmology, School of Life and Medical Sciences (SLMS) and only accessible to members of the AlzEye research team. All data entities were listed within the UCL SLMS Information Asset Register.

Patient and public involvement and engagement

Patient and public involvement and engagement (PPIE) support was provided by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at Moorfields Eye Hospital and UCL Institute of Ophthalmology and has been embedded throughout this project from priority setting to plans for dissemination. Feedback has been sought through public engagement events, survey of eye service users and reports within the media. Patients and the public actively contributed to identifying the priority setting of dementia and the acceptability of using routinely acquired eye scans for research purposes and without consent. In addition, two members of the public will sit on the AlzEye working group to contribute to results interpretation and co-authoring and dissemination of research outputs. The members will be supported in selecting the results they find relevant and presenting them to wider patient communities.

Ophthalmic health variables

Patient-level ophthalmic variables were extracted from the MEH data warehouse, which aggregates information from the patient administration system (PAS), electronic health record and imaging database, all linked through a unique MEH hospital identification number. Sociodemographic data, including date of birth, sex, ethnicity and postcode as well as patient's clinic appointments and operation dates are housed within PAS. Surgical procedures were recorded in the electronic health record (EHR) at MEH from September 4th, 2012. Operation details, including procedure name, laterality and indication for surgery are contained within the MEH EHR and uploaded to the MEH data warehouse. A patient undergoing the most common operation in the United Kingdom, cataract extraction, would therefore have an entry for the typical procedure, (*phacoemulsification and intraocular lens implant*) operated eye (*right or left*) and indication (*cataract*).

Colour retinal photography (Figure 2A) and optical coherence tomography (OCT, Figure 2B) images, which represent the majority of retinal images within the database, have been processed through segmentation and feature extraction software. The Vascular Assessment and Measurement Platform for Images of the Retina (VAMPIRE) system provides fully automated segmentation and extraction of retinal fractal dimensions, a measure of the geometric complexity of retinal vasculature, from colour retinal photographs with future plans to derive retinal vascular calibre[22,23]. OCT scans are segmented and retinal sublayer thicknesses computed using the Topcon Advanced Biomedical Imaging Laboratory (TABIL) software[24]. Retinal sublayer analyses will conform to the consensus nomenclature outlined by the International Nomenclature for Optical Coherence Tomography Panel (Figure 2E)[25].

For the purposes of this report, four common ophthalmic diseases were described - cataract, glaucoma, neovascular age-related macular degeneration (AMD) and proliferative diabetic retinopathy (PDR).

Cataract was defined as any operation code denoting phacoemulsification surgery and the indication of cataract. For the purposes of this report, only first eye cataract surgery was included.

Glaucoma was defined as any patient attending the glaucoma clinic three or more times with ongoing follow up from January 1st 2010. The first two years of the study period were excluded as this may have incorporated patients with previous diagnoses of glaucoma where the maximum follow up interval can approach two years; in contrast any patient being seen after 2 years since study inception with no previous visit within that 2 year period can be assumed to have/carry a new diagnosis of glaucoma.

Diabetic eye disease represents a special case due to audit procedures mandated by the NHS Diabetic Eye Screening Programme. Coding of eye disease secondary to diabetes mellitus is rigorously validated by a dedicated team within MEH with a dedicated database consisting of all patients with diabetic eye disease and their retinopathy and maculopathy grades according to the NHS Diabetic Eye Screening Programme criteria[26], at hospital appointment from September 12, 2013 onwards. Dates for onset of proliferative diabetic retinopathy dates were recorded as the first appointment for each patient where this diagnosis was first made.

AMD can be categorised into two major types - dry and neovascular (“wet”). Dry AMD is slowly progressive and with no active hospital intervention currently available; it is thus MEH standard practice for patients with features typical of dry AMD to be discharged with lifestyle and monitoring advice (self-monitoring and standard optometric review). In contrast, neovascular AMD requires treatment through intravitreal anti-VEGF injections, and therefore remains under active follow-up. The diagnostic codes for neovascular AMD were based on extensive previous

work in which all patients with neovascular AMD at MEH were manually validated up to 2018[27,28].

Systemic health variables

Systemic health data were derived from HES APC data, with a focus on cardiovascular disease and all-cause dementia. Diagnostic codes in HES APC are reported in line with the 10th revision of ICD[29]. In line with previous reports, myocardial infarction was defined as code I21 or I22[30–32]. Similarly, stroke was defined using stroke definitions from UK Biobank[33]. Dementia was defined as ICD codes E512 (Wernicke's encephalopathy), F00 (Dementia in Alzheimer disease), F01 (Vascular dementia), F02 (Dementia in other diseases classified elsewhere), F03 (Unspecified dementia), F10.6 (Mental and behavioural disorders due to psychoactive substance use, Amnesic syndrome) F10.7 (Mental and behavioural disorders due to psychoactive substance use, Residual and late-onset psychotic disorder) , G30 (Alzheimer disease), or G31.0 (Other degenerative diseases of nervous system, not elsewhere classified), derived from previous work evaluating the agreement between HES admitted patient care data and primary care data, through general practitioner surveys and the Clinical Practice Research Datalink (CPRD)[34]. The unique number of patients specifically with Alzheimer's disease (F00, G30) and vascular dementia (F01) are also described.

Data Linkage and Transfer

The linkage strategy was designed through collaboration between experts in information governance, information technology, computer scientists and clinicians based at MEH, University College London (UCL) and NHS Digital (Figure 3). A third-party linkage approach was used for two main reasons. Firstly, it enhanced privacy-preservation as the data originator,

MEH, never received HES admissions data and the third party, UCL, did not receive personally identifiable information. Secondly, it enabled the linked dataset to be accessible within a site with sufficient high-performance computing capability to undertake the proposed analyses, a function significantly beyond almost all NHS facilities. Patient link identifiers consisting of a unique NHS identification number, sex and date of birth originating from MEH were transferred to NHS Digital in conjunction with a unique study ID generated using a cryptographic hash function (random pseudonmisation). Ophthalmic covariates, mortality data, and patient sociodemographics with study ID were transferred to UCL. Ophthalmic imaging data pertaining to the patients within the study were extracted and de-identified during conversion from their proprietary format to Digital Imaging and Communications in Medicine (DICOM) format before transfer to UCL. Following linkage with HES, NHS Digital transferred HES data to the UCL Data Safe Haven, a “walled garden” trusted research environment certified and externally audited to ISO27001 information security standards[35,36].

Statistical analysis

Imaging-based studies within the AlzEye study are generally planned to take the form of nested case-control studies. To improve efficiency, controls may be matched with cases by age and sex, using conditional logistic regression for statistical modelling of binary outcomes and survival analysis for time-to-event data (e.g. Cox Proportional hazard modelling)[37]. In cases where the competing risk of death is prominent, subdistribution hazard ratios with 95% confidence intervals will be estimated as a sensitivity analysis using the multivariable semiparametric competing risks proportional hazards model approach described by Fine and Gray[38]. Alternative high-dimensional modelling approaches, such as convolutional neural networks and vision transformers, will also be explored. Prior to receipt of HES data from NHS Digital, sample size calculations were undertaken. Specifically, we evaluated the association

between OCT-derived peripapillary retinal nerve fibre layer and macular ganglion cell-inner plexiform layer thicknesses and dementia. Given an odds ratio of 1.4 with an alpha of 5% and a power of 90% on a 1:1 matched study design, a total sample size of 2106 is required[39].

Figures for this report were designed in R version 4.1.0 (R Core Team, 2021. R Foundation for Statistical Computing, Vienna, Austria).

FINDINGS TO DATE

Extraction of unique patients attending MEH outpatient clinics between January 1st 2008 and April 1st 2018 generated a cohort of 353,191 unique patients. Thirty-four patients with invalid or missing linking variables were excluded leaving a total of 353,157. A breakdown of sociodemographic details by category of the cohort are provided in Table 1. Of the cohort, 190,494 were female (53.9%) and the mean age was 68.4 (+/- 13.9) years. Among those who self-reported ethnicity (n=271,293, 76.8%), 48,119 were South Asian, 31,614 Black, 2966 Mixed, 52,851 Other ethnic group and 135,743 White. Of the 353,157 patients, 186,651 had a total of 1,337,711 HES episodes in the study period. NHS Digital performs a hierarchical stepwise linkage approach providing a 'Match Rank' for each HES episode[40]. Among the 1,337,711 HES episodes matched, Match Rank was two for 1,337,482 episodes (exact NHS number, exact date of birth and exact sex linked), four for 46 episodes (exact NHS number, exact sex and partial date of birth), and eight for 183 episodes (exact NHS number).

An illustration of the major common ophthalmic diseases within the cohort is shown in a CONSORT-style diagram in Figure 4. Following the case definition, and exclusion of invalid dates, a total of 59,102 patients had first eye cataract surgery, 31,060 glaucoma, 7214 neovascular AMD and 2494 PDR.

	Characteristic	N (%)
	All	353,157
Sex	**Female**	190,494 (53.9)
	Male	162,663 (46.1)
Age group*	**40-49 years**	35,262 (10.0)
	50-59 years	66,101 (18.7)
	60-69 years	79,018 (22.4)
	70-79 years	84,942 (24.1)
	80+ years	87,834 (24.9)
Ethnicity	**Black**	31,614 (9.0)
	White	135,743 (38.4)
	South Asian	48,119 (13.6)
	Other/Unknown	137,681 (39.0)
Index of multiple deprivation decile	**1 (most deprived)**	18,194 (5.2)
	2	50,443 (14.3)
	3	50,869 (14.4)
	4	42,603 (12.1)
	5	38,964 (11.0)
	6	36,906 (10.5)
	7	31,317 (8.9)
	8	28,180 (8.0)
	9	29,906 (8.5)
	10 (least deprived)	24,610 (7.0)
	Unknown	1165 (0.3)

Table 1: Baseline sociodemographic characteristics of the AlzEye cohort. Data are shown as n(%).

*Age is taken as that of April 1st, 2018.

Among the 187,811 patients with recorded HES episodes, 12,022 patients had episodes with coded myocardial infarction, 11,735 patients with all-cause stroke and 13,363 with dementia. Within the dementia group, 4487 patients had codes that were specific for Alzheimer's dementia and 3381 for vascular dementia (Table 2).

Group	Disease	ICD code(s)	Number of patients
Cardiovascular	Acute coronary syndrome	I21, I22	12,022
	Heart failure	I50	24,034
	Atrial Fibrillation	I48	32,848
	Hypertension	I10, I15	151,937
	Subarachnoid haemorrhage	I60	642
	Intracerebral hemorrhage	I61	1865
	Ischaemic stroke	I63-I64	9996
	All stroke	I60, I61, I63, I64	11,735
Neurodegenerative	Alzheimer's disease	F00, G30	4487
	Vascular dementia	F01	3381
	Parkinson's disease	G20	3211
	All-cause dementia	E12, F00, F01, F02, F03, F106, F107, G30, G310	13,363
Other	Diabetes mellitus (Type 1 and 2)	E10, E11	71,570

Table 2: Number of patients by selected examples of specified 10th revision of International Classification of Diseases (ICD) codes relating to diabetes mellitus, cardiovascular and neurodegenerative diseases.

Imaging

During the study period, a total of 6,261,931 images were acquired from 154,830 patients. The three leading image modalities were colour retinal photographs (n=1,874,175), OCT (n=1,567,358) and red-free photographs (n=1,147,487). The distribution of imaging modalities across the three vendors used for retinal imaging at MEH - Topcon (Topcon corp, Tokyo, Japan), Heidelberg (Heidelberg Engineering, Heidelberg, Germany), and Optos (Optos,

Dunfermline, UK) - are shown in Figure 5 and Table 2. Most images were acquired on the Topcon system (n=5,553,826, 88.7%). Number of images by year is shown in Figure 6. During the study period, annual imaging acquisition increased from 229,868 scans in 2008 to 1,021,904 in 2017. For 2018, collection stopped on April 1st precluding a complete annual figure.

Vendor	Modality	Number of images	Number of patients
Topcon	Angiography	1,128,723	21,225
	Autofluorescence	11,761	2078
	Colour photography	1,874,175	139,307
	Red-free	1,146,854	122,453
	OCT	1,391,826	138,911
	Other	487	48
Heidelberg	Angiography	89,264	4061
	Autofluorescence	94,533	16,863
	Infrared	192,634	21,676
	OCT	175,532	21,191
	Other	19,781	2439
Optos	Angiography	77,813	2215
	Autofluorescence	18,590	5666
	Pseudocolour photography	39,958	6887

Table 3: Retinal imaging within the AlzEye dataset by vendor and imaging modality. OCT: Optical coherence tomography. Angiography refers to dye-based techniques (fluorescein and indocyanine green).

STRENGTHS AND LIMITATIONS

To our knowledge, we have created the world's largest retinal imaging research dataset available presently, linking secondary healthcare ophthalmic data from 353,157 patients seen over a ten-year period with information on general health and key systemic diseases, as captured through admissions to any hospital within the NHS of England. This comprises 6,261,931 images, obtained using seven different modalities from three different manufacturers, in 154,830 patients. The current large-scale UK health database, UKBB, provides useful context for AlzEye. Cross-sectional data are available in UKBB with two retinal imaging modalities (colour retinal photography and OCT) obtained using technology from one manufacturer (Topcon), and at a single time point in 67,321 people. Notwithstanding the recognised limitations (see subsection 'Limitations of the AlzEye cohort') of real-world datasets and the coding within the HES database, AlzEye provides a number of distinct advantages beyond purely scale. Retinal imaging data are longitudinal, highly multimodal, and pertain to an ethnically and socioeconomically diverse cohort representative of the adult population with eye disease. Moreover, AlzEye has demonstrated relatively low cost. The study is funded through a charity small grant award and NIHR Biomedical Research Centre support amounting to £20,000.

Comparison with other resources

UKBB is the major comparator for AlzEye, being the largest of the prospective epidemiological cohort datasets which provide cross-sectional retinal imaging in association with systemic disease variables[41]. One of the limitations of UKBB is that, unlike AlzEye, it does not provide longitudinal retinal images. Another prospective epidemiological cohort study, the Rotterdam study, does collect longitudinal retinal imaging data from approximately 15,000 participants, of which 5065 participants were eligible for OCT scanning as of September 2017[42]. The

Rotterdam Study has uncovered several landmark findings, particularly in regards to causal determinants, but its cohort remains relatively small in comparison to UKBB and AlzEye with the majority of participants recruited from one district within Rotterdam, Netherlands[7,10]. The Singapore Epidemiology of Eye Disease (SEED), is one longitudinal multimodal retinal imaging initiative which is underway, in which 10,033 participants of Chinese, Indian and Malay ethnicity have been recruited to undergo six-yearly retinal imaging[43]. Whilst the importance of such initiatives is often highlighted, this does not appear to have yet translated into actual research-ready resources. A recent review of ophthalmic imaging datasets did not reveal any additional relevant publicly available datasets that included linked systemic health data[44]. Additionally, our own review of the literature has not identified any examples of large-scale linked real-world datasets (i.e. including those with restricted access) which include linked systemic health data. The scarcity of such resources suggests that the construction of such datasets is challenging to undertake, presumably due to factors such as cost, required duration and delayed output, retention of participants, and concerns over technological redundancy. The AlzEye approach is an important alternative model in this context.

Potential research impact from the novel AlzEye cohort

Several epidemiological opportunities arise with AlzEye. Firstly, it provides a real-world snapshot of ophthalmic secondary care use, representing approximately 1.2% of the UK population aged 40 and above (27,858,459 in 2011). This is a powerful tool for informing public health and policymaking in eye services[45], and is exceptional in characterising the potential impact that may arise from the intersect between disabling diseases such as stroke and proliferative diabetic retinopathy.

Secondly, it allows the identification and exploration of relationships between newly diagnosed ophthalmic disease (or newly referred to hospital eye services) and emerging systemic events

and accruing multimorbidity. Patients tend to respond early to issues with their sight and an understanding of how an ophthalmic presentation is linked to an increased likelihood of serious systemic disease may provide an opportunity for earlier intervention in those diseases[46].

Thirdly, nested case-control studies evaluating retinal-based oculomic biomarkers in those with systemic diseases (e.g. dementia) can provide insight into their value in either static or dynamic risk prediction. Newer modelling approaches have highlighted the potential utility of the retina in screening for and risk stratification of cardiovascular, neurodegenerative, renal, hepatic and haematological diseases[47–52];[53].

Finally, by its magnitude and wealth of high-quality labels, both ophthalmic and systemic, AlzEye provides a powerful catalyst for high-dimensional model development, echoing that of Imagenet, a database currently exceeding 14 million images, which propelled deep learning and computer vision research forward a decade ago[54].

Lessons learned from the AlzEye approach

AlzEye highlights an opportunity for maximising the value of routinely-collected data to support research for patient benefit. However, there are a number of governance and technical challenges when undertaking large scale investigator-led data linkage[55]. In AlzEye, early cross-disciplinary dialogue between experts in information governance, information technology and data protection at both institutional parties (MEH and UCL) as well as the data production team at NHS Digital established a privacy-by-design linkage approach, which enhanced privacy preservation while maintaining computational feasibility[56][57]. At its worst, an intrusion of the identifiable data during the development of AlzEye would have informed the violator that a given individual had visited MEH at some point between 2008 and 2018. Due to the novel approach of AlzEye within our centre, the greatest governance hurdle was securing study sponsorship, a

process which took nearly eight months. Once approved, permissions from the bodies of the Health Research Authority, and overall approval were given within eight weeks. Linking with high-dimensional imaging data also posed a number of technical obstacles. As highlighted recently by the American Academy of Ophthalmology (AAO), ophthalmic imaging technologies suffer from limited interoperability and low compliance to standardised formats, such as DICOM, the AAO also commenting that 'even so-called DICOM-compliant devices fail to meet DICOM standards with significant limitations, such as the embedding of patient identifiers on the image'[58]. A key undertaking within AlzEye was thus the secure and robust but efficient fully-automated processing of raw ophthalmic imaging data from its proprietary file format with associated personal metadata to standard DICOM form with the identifiers stripped. Fortunately, while this operation requires significant technical and engineering input, most medical imaging modalities already benefit from standardisation among vendors obviating this step for other researchers seeking to emulate our approach. Finally, a key objective of AlzEye is the development of clinical prediction models using deep learning approaches, which require significant computing capacity. Provisions for four graphics processing units (GPU) housed within UCL enable this step however other centres may additionally consider recent guidance on the safeguards required for locating health data within cloud environments and the implications this brings for accessing virtual GPUs[59].

Limitations of the AlzEye cohort

Despite the opportunities afforded by AlzEye, there are a number of limitations to this kind of approach and potential sources of bias. Firstly, caution must be paid to the validity of HES diagnostic coding[60]. Although previous validation studies have concluded that discharge coding within HES is sufficiently robust for research purposes[34,61], sizable proportions of cases may be missed when using individual sources[62]. For example, recent work linking the electronic health records of 54.4 million people in England showed that HES captured 80.5%

and 65% of myocardial infarctions and stroke/transient ischaemic attacks respectively when compared to linkage additionally incorporating Death Registry and primary care records[63]. One mitigation strategy for this source of bias for real world data is therefore linking to multiple sources. In terms of selection bias, as a hospital-attending cohort, the individuals within the AlzEye cohort are likely to have greater medical comorbidity than the general population, limiting the external validity of any findings. In addition, by the very nature of the dataset, patients within the AlzEye cohort will have definite or suspected ophthalmic disease, particularly among those with repeated retinal imaging. The risk of under-recording of potentially important variables such as smoking may also lead to residual confounding.

The enrichment of multimodal health data acquired as part of a patient's routine clinical care with nationally-held databases provides a powerful foundation for discovery science and epidemiological research. We highlight key considerations and challenges for those seeking to link high-dimensional data sources, from high-resolution imaging to waveform data, with locally-held specialist data. Additionally, we provide the cohort profile for AlzEye, a powerful platform for oculomic discovery, specifically evaluating the association between retinal morphology, and both cardiovascular diseases and dementia. Beyond discovery, the ALlzEye cohort is anticipated to become an important resource for the development and validation of deep learning-based clinical prediction models that may enable earlier intervention for patients at risk of these life-threatening conditions.

LEGENDS

Figure 1: Schematic of the key milestones, prerequisites and approvals with their corresponding achievement dates for the AlzEye dataset. REC: Research Ethics Committee, CAG: Confidential Advisory Group, HRA: Health Research Authority, NHS: National Health Service, IGARD: Independent Group Advising on the Release of Data, DSA: Data Sharing Agreement.

Figure 2: Composite figure showing the major retinal imaging modalities within AlzEye. A: Colour fundus photograph, B: Red-free photograph, C: Fundus autofluorescence (widefield), D: Pseudocolour photography (widefield) and E: Optical coherence tomography of the central macula illustrating segmentation of the individual sublayers. Consensus nomenclature for the retinal sublayers is indicated. NFL= nerve fibre layer, GCL = ganglion cell layer, IPL: inner plexiform layer, INL: inner nuclear layer, OPL: outer plexiform layer, ONL = outer nuclear layer, MZ = myoid zone, RPE-IZ = RPE-interdigitation zone, BM = Bruch's membrane.

Figure 3: Linkage approach of AlzEye. Moorfields Eye Hospital NHS Foundation Trust securely transfers a spreadsheet of identifiers with a study ID to NHS Digital and separately transfers the study ID with ophthalmic data, including diagnoses and retinal images, to University College London. NHS Digital links the identifiers with the Hospital Episode Statistics (HES) database and returns the admissions data with the study ID (and no identifiable data) to UCL. UCL links the ophthalmic data from Moorfields Eye Hospital with HES data from NHS Digital using the study ID.

Figure 4: CONSORT style flow chart illustrating the distribution of cataract, glaucoma, neovascular age-related macular degeneration (AMD) and proliferative diabetic retinopathy (PDR) within the AlzEye dataset.

Figure 5: Parallel sets diagram illustrating the imaging modality across vendors within AlzEye. The majority of images were acquired on the Topcon system and the most frequent modalities were colour photography and optical coherence tomography. Designed using the `networkD3` package.

Figure 6: Stacked bar chart of the annual number of images acquired during the study period for the three leading device vendors at Moorfields Eye Hospital. Data for 2018 represents 3 months only prior to the study end-date.

ACKNOWLEDGEMENTS

We thank Karen Bonstein and Andi Skilton for support with patient and public involvement and engagement, Menachem Katz, Ben Ward, Ross Green, Maxim Daniline and Simon St John-Green for information technology support, Anthony Peacock for advice on use of Trusted Research Environments and Antonio de la Plaza Larrea and Richard Macmillan for legal guidance on data sharing agreements. We also are grateful to Anthony Khawaja and Cathie Sudlow for feedback on study design.

CONTRIBUTORS

SW, AD and PK wrote the first draft of the manuscript, which was critically revised by FH, MCB, RS, NP, XL, HM, DA, ET, SP, KB, JH, AP, and JR. SW, FH, NP, HM, JH, AP, JR, AD and PK were involved in the original design of the study. Computer science expertise was through RS, NP and DA, information governance through JH. MCB, AP, JR and AD provided statistical and epidemiological guidance. All authors have approved the final version of this manuscript.

FUNDING

The study was funded through a small grant awarded by Fight for Sight and Alzheimer's Research UK. Infrastructural support was through the NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology. The funders had no role in the conception, design or analysis of the study.

COLLABORATION

National and international collaborations are welcomed though restrictions on access to the cohort mean that only the AlzEye researchers can directly analyse individual-level systemic health data. Interested researchers should contact s.wagner@ucl.ac.uk.

DATA AVAILABILITY STATEMENT

The data are subject to the contractual restrictions of the DSA between NHS Digital, Moorfields Eye Hospital and University College London and are therefore not available for access beyond the AlzEye research team.

COMPETING INTERESTS STATEMENT

SW is funded through a Medical Research Council Clinical Research Training Fellowship (MR/TR000953/1).

FH: None.

MCB: None.

NP is funded by a Moorfields Eye Charity Career Development Award (R190031A).

RS: None.

XL: None.

HM is supported by the National Institute of Health Research's Comprehensive Biomedical Research Centre (BRC) at University College London Hospitals (UCLH).

DA: None

ET: None

SP receives support from the National Institute for Health Research Biomedical Research Centre at Barts.

KB has received speaker fees from Novartis, Bayer, Alimera, Allergan, Roche, and Heidelberg; meeting or travel fees from Novartis and Bayer; compensation for being on an advisory board from Novartis and Bayer; consulting fees from Novartis and Roche; and research support from Apellis, Novartis, and Bayer.

JH: None

AP receives financial support from the National Institute for Health Research Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. AP is part of the steering committee of the ANGI network which is sponsored by ZEISS, steering committee of the OCTiMS study which is sponsored by Novartis, and reports speaker fees from Heidelberg Engineering.

JR receives support from the National Institute for Health Research as a Senior Investigator and via the National Institute for Health Research Biomedical Research Centres at Moorfields Eye Hospital and Great Ormond Street Hospital.

AD is Director of INSIGHT, the HDRUK Health Data Research Hub for Eye Health.

PK is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1); receives research support from Apellis; is a consultant for DeepMind, Roche, Novartis, Apellis, and BitFount; is an equity owner in Big Picture Medical; and has received speaker fees from Heidelberg Engineering, Topcon, Allergan, Roche, and Bayer; meeting or travel fees from Novartis and Bayer; and compensation for being on an advisory board from Novartis and Baye

REFERENCES

1 Munevar S. Unlocking Big Data for better health. *Nat Biotechnol* 2017;**35**:684–6.

2 Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* 2020;**26**:29–38.

3 Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;**375**:1216–9.

4 Bhardwaj R, Sethi A, Nambiar R. Big data in genomics: An overview. In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE 2014. doi:10.1109/bigdata.2014.7004392

5 He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. *Int J Mol Sci* 2017;**18**. doi:10.3390/ijms18020412

6 Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;**278**:563–77.

7 Hofman A, Breteler MMB, van Duijn CM, *et al.* The Rotterdam Study: objectives and design update. *Eur J Epidemiol* 2007;**22**:819–29.

8 Riboli E, Hunt KJ, Slimani N, *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;**5**:1113–24.

9 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9.

10 Ikram MA, Brusselle GGO, Murad SD, *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* 2017;**32**:807–50.

11 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.

12 Chaudhry Z, Mannan F, Gibson-White A, *et al.* Research Outputs of England's Hospital Episode Statistics (HES) Database: Bibliometric Analysis. Journal of Innovation in Health Informatics. 2017;**24**:329. doi:10.14236/jhi.v24i4.949

13 Zhu Y, Matsuyama Y, Ohashi Y, *et al.* When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *J Biomed Inform* 2015;**56**:80–6.

14 Hospital information boosts UK Biobank resource. https://www.ukbiobank.ac.uk/2013/09/20000-participants-return-for-a-repeat-assessment/ (accessed 13 Oct 2020).

15 Boyd A, Golding J, Macleod J, *et al.* Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. International Journal of Epidemiology. 2013;**42**:111–27. doi:10.1093/ije/dys064

16 Luben R, Hayat S, Khawaja A, *et al.* Residential area deprivation and risk of subsequent hospital admission in a British population: the EPIC-Norfolk cohort. BMJ Open. 2019;**9**:e031251. doi:10.1136/bmjopen-2019-031251

17 Large P. Annual mid-year population estimates, UK - Office for National Statistics. 2014.https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/2014-06-26 (accessed 23 Jul 2021).

18 Ministry of Housing, Communities & Local Government. English indices of deprivation 2015. 2015.https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 (accessed 18 Sep 2020).

19 Great Britain. *National Health Service Act 2006*. The Stationery Office 2006.

20 Data Access Request Service (DARS). https://digital.nhs.uk/services/data-access-request-service-dars (accessed 7 Jul 2021).

21 Independent Group Advising on the Release of Data. https://webarchive.nationalarchives.gov.uk/20200706184649/https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data (accessed 7 Jul 2021).

22 Perez-Rovira A, MacGillivray T, Trucco E, *et al.* VAMPIRE: Vessel assessment and measurement platform for images of the REtina. *Conf Proc IEEE Eng Med Biol Soc* 2011;**2011**:3391–4.

23 Liew G, Wang JJ, Cheung N, *et al.* The retinal vasculature as a fractal: methodology, reliability, and relationship to blood pressure. *Ophthalmology* 2008;**115**:1951–6.

24 Keane PA, Grossi CM, Foster PJ, *et al.* Optical Coherence Tomography in the UK Biobank Study - Rapid Automated Analysis of Retinal Thickness for Large Population-Based Studies. *PLoS One* 2016;**11**:e0164095.

25 Staurenghi G, Sadda S, Chakravarthy U, *et al.* Proposed lexicon for anatomic landmarks in normal posterior segment spectral-domain optical coherence tomography: the IN•OCT consensus. *Ophthalmology* 2014;**121**:1572–8.

26 Peate I. The NHS diabetic eye screening programme. British Journal of Healthcare Assistants. 2019;**13**:596–9. doi:10.12968/bjha.2019.13.12.596

27 Fasler K, Fu DJ, Moraes G, *et al.* Moorfields AMD database report 2: fellow eye involvement with neovascular age-related macular degeneration. *Br J Ophthalmol* 2020;**104**:684–90.

28 Fasler K, Moraes G, Wagner S, *et al.* One- and two-year visual outcomes from the Moorfields age-related macular degeneration database: a retrospective cohort study and an open science resource. *BMJ Open* 2019;**9**:e027441.

29 ICD-10 Version:2010. https://icd.who.int/browse10/2010/en (accessed 22 Jul 2021).

30 Asaria P, Elliott P, Douglass M, *et al.* Acute myocardial infarction hospital admissions and deaths in England: a national follow-back and follow-forward record-linkage study. The

Lancet Public Health. 2017;**2**:e191–201. doi:10.1016/s2468-2667(17)30032-4

31 Metcalfe A, Neudam A, Forde S, *et al.* Case Definitions for Acute Myocardial Infarction in Administrative Databases and Their Impact on In-Hospital Mortality Rates. Health Services Research. 2013;**48**:290–318. doi:10.1111/j.1475-6773.2012.01440.x

32 McCormick N, Lacaille D, Bhole V, *et al.* Validity of myocardial infarction diagnoses in administrative databases: a systematic review. *PLoS One* 2014;**9**:e92286.

33 [No title]. https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/alg_outcome_stroke.pdf (accessed 23 Aug 2021).

34 Brown A, Kirichek O, Balkwill A, *et al.* Comparison of dementia recorded in routinely collected hospital admission data in England with dementia recorded in primary care. *Emerg Themes Epidemiol* 2016;**13**:11.

35 Lea NC, Nicholls J, Dobbs C, *et al.* Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. JMIR Medical Informatics. 2016;**4**:e22. doi:10.2196/medinform.5571

36 Certificate Client Directory search results. https://www.bsigroup.com/en-GB/our-services/certification/certificate-and-client-directory/search-results/?searchkey=licence%3dIS%2b612909%26company%3duniversity%2bcollege%2bLondon&licencenumber=IS%20612909 (accessed 26 Jul 2021).

37 Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;**34**:187–202. doi:10.1111/j.2517-6161.1972.tb00899.x

38 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;**94**:496–509. doi:10.1080/01621459.1999.10474144

39 Mutlu U, Colijn JM, Ikram MA, *et al.* Association of Retinal Neurodegeneration on Optical Coherence Tomography With Dementia: A Population-Based Study. *JAMA Neurol* 2018;**75**:1256–63.

40 Harper G. Linkage of Maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005–2014: Quality assurance of linkage of routine data for singleton and multiple births. BMJ Open. 2018;**8**:e017898. doi:10.1136/bmjopen-2017-017898

41 Chua SYL, Thomas D, Allen N, *et al.* Cohort profile: design and methods in the eye and vision consortium of UK Biobank. *BMJ Open* 2019;**9**:e025077.

42 Mutlu U, Bonnemaijer PWM, Ikram MA, *et al.* Retinal neurodegeneration and brain MRI markers: the Rotterdam Study. *Neurobiol Aging* 2017;**60**:183–91.

43 Majithia S, Tham Y-C, Chee M-L, *et al.* Cohort Profile: The Singapore Epidemiology of Eye Diseases study (SEED). *Int J Epidemiol* 2021;**50**:41–52.

44 Khan SM, Liu X, Nath S, *et al.* A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021;**3**:e51–66.

45 Age groups. 2018.https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest#:~:text=by%20ethnicity%20Summary-,The%20data%20shows%20that%3A,aged%2060%20years%20and%20over (accessed 15 Jul 2021).

46 Enoch J, McDonald L, Jones L, *et al.* Evaluating Whether Sight Is the Most Valued Sense. *JAMA Ophthalmol* 2019;**137**:1317–20.

47 Sabanayagam C, Xu D, Ting DSW, *et al.* A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *Lancet Digit Health* 2020;**2**:e295–302.

48 Mitani A, Huang A, Venugopalan S, *et al.* Detection of anaemia from retinal fundus images via deep learning. *Nat Biomed Eng* 2020;**4**:18–27.

49 Poplin R, Varadarajan AV, Blumer K, *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;**2**:158–64.

50 Wisely CE, Wang D, Henao R, *et al.* Convolutional neural network to identify symptomatic Alzheimer's disease using multimodal retinal imaging. *Br J Ophthalmol* Published Online First: 26 November 2020. doi:10.1136/bjophthalmol-2020-317659

51 Cheung CY, Xu D, Cheng C-Y, *et al.* A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nat Biomed Eng* 2021;**5**:498–508.

52 Xiao W, Huang X, Wang JH, *et al.* Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. *Lancet Digit Health* 2021;**3**:e88–97.

53 Wagner SK, Fu DJ, Faes L, *et al.* Insights into Systemic Disease through Retinal Imaging-Based Oculomics. *Transl Vis Sci Technol* 2020;**9**:6.

54 Deng J, Dong W, Socher R, *et al.* ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. doi:10.1109/cvprw.2009.5206848

55 Harron K, Dibben C, Boyd J, *et al.* Challenges in administrative data linkage for research. Big Data & Society. 2017;**4**:205395171774567. doi:10.1177/2053951717745678

56 Data protection by design and default. Published Online First: 9 February 2021.https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/ (accessed 13 Jul 2021).

57 [No title]. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf (accessed 13 Jul 2021).

58 Lee AY, Campbell JP, Hwang TS, *et al.* Recommendations for Standardization of Images in Ophthalmology. *Ophthalmology* 2021;**128**:969–70.

59 NHS and social care data: off-shoring and the use of public cloud services. https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-

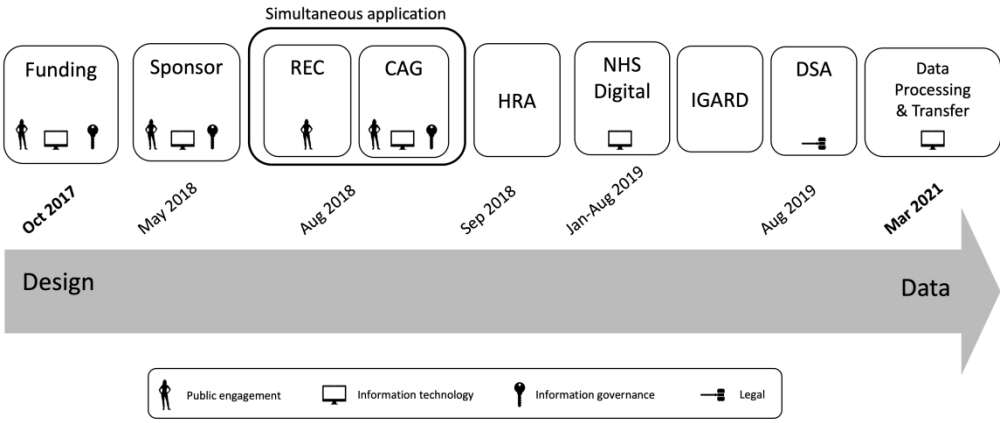
information-governance/nhs-and-social-care-data-off-shoring-and-the-use-of-public-cloud-services (accessed 13 Jul 2021).

60 Sinha S, Peach G, Poloniecki JD, *et al.* Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes—systematic review and recommendations for reporting. European Journal of Public Health. 2013;**23**:86–92. doi:10.1093/eurpub/cks046

61 Burns EM, Rigby E, Mamidanna R, *et al.* Systematic review of discharge coding accuracy. *J Public Health* 2012;**34**:138–48.

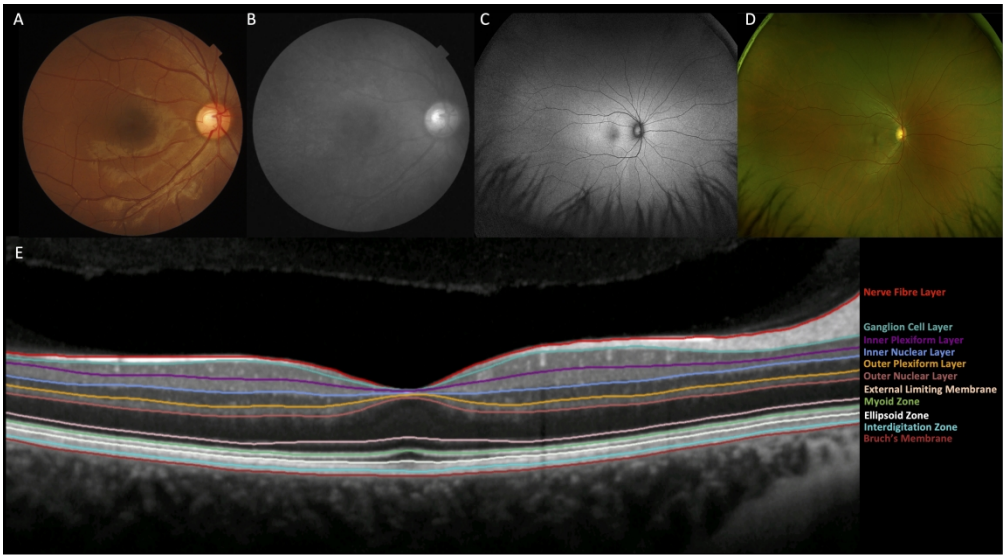
62 Herrett E, Shah AD, Boggon R, *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;**346**:f2350.

63 Wood A, Denholm R, Hollings S, *et al.* Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021;**373**:n826.



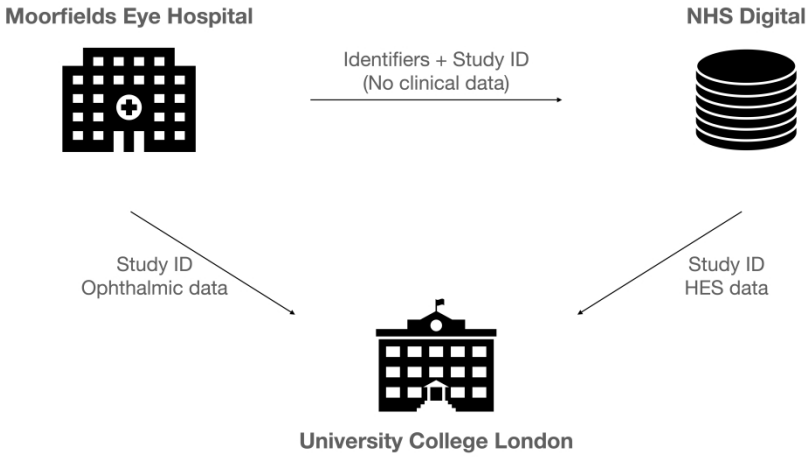
Schematic of the key milestones, prerequisites and approvals with their corresponding achievement dates for the AlzEye dataset. REC: Research Ethics Committee, CAG: Confidential Advisory Group, HRA: Health Research Authority, NHS: National Health Service, IGARD: Independent Group Advising on the Release of Data, DSA: Data Sharing Agreement.

1327x627mm (72 x 72 DPI)



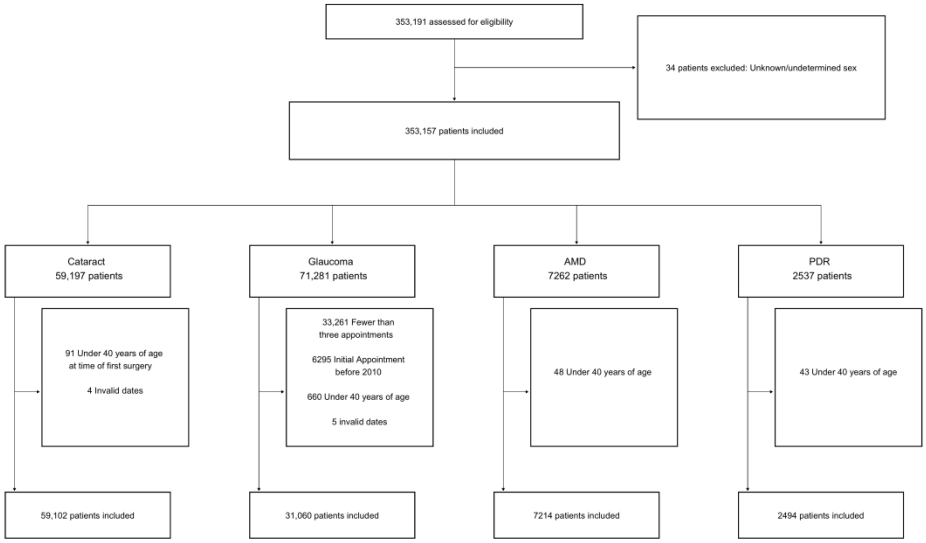
Composite figure showing the major retinal imaging modalities within AlzEye. A: Colour fundus photograph, B: Red-free photograph, C: Fundus autofluorescence (widefield), D: Pseudocolour photography (widefield) and E: Optical coherence tomography of the central macula illustrating segmentation of the individual sublayers. Consensus nomenclature for the retinal sublayers is indicated. NFL= nerve fibre layer, GCL = ganglion cell layer, IPL: inner plexiform layer, INL: inner nuclear layer, OPL: outer plexiform layer, ONL = outer nuclear layer, MZ = myoid zone, RPE-IZ = RPE-interdigitation zone, BM = Bruch's membrane.

1456x803mm (72 x 72 DPI)



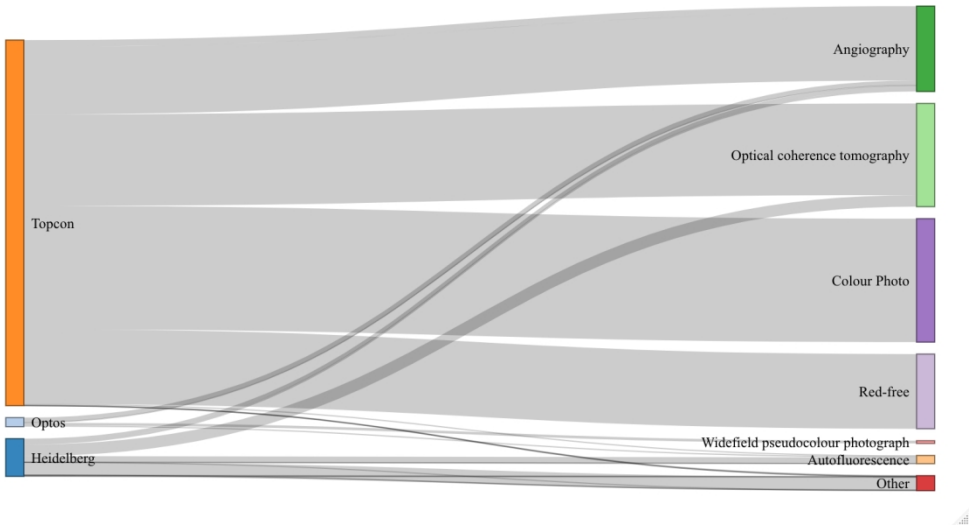
Linkage approach of AlzEye. Moorfields Eye Hospital NHS Foundation Trust securely transfers a spreadsheet of identifiers with a study ID to NHS Digital and separately transfers the study ID with ophthalmic data, including diagnoses and retinal images, to University College London. NHS Digital links the identifiers with the Hospital Episode Statistics (HES) database and returns the admissions data with the study ID (and no identifiable data) to UCL. UCL links the ophthalmic data from Moorfields Eye Hospital with HES data from NHS Digital using the study ID.

2822x1587mm (72 x 72 DPI)



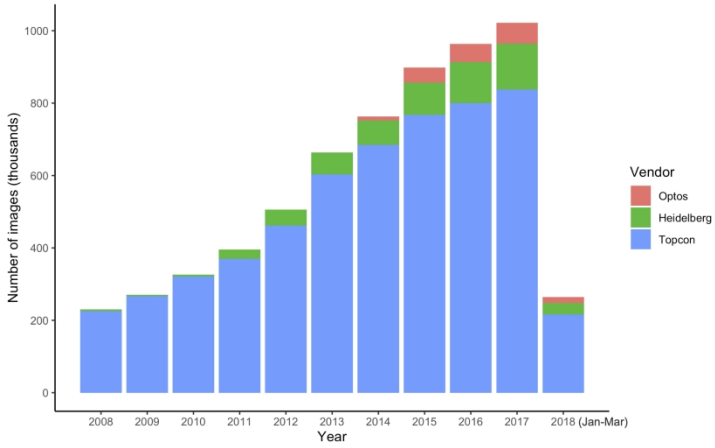
CONSORT style flow chart illustrating the distribution of cataract, glaucoma, neovascular age-related macular degeneration (AMD) and proliferative diabetic retinopathy (PDR) within the AlzEye dataset.

1693x1058mm (72 x 72 DPI)



Parallel sets diagram illustrating the imaging modality across vendors within AlzEye. The majority of images were acquired on the Topcon system and the most frequent modalities were colour photography and optical coherence tomography. Designed using the networkD3 package.

584x324mm (72 x 72 DPI)



Stacked bar chart of the annual number of images acquired during the study period for the three leading device vendors at Moorfields Eye Hospital. Data for 2018 represents 3 months only prior to the study end-date

1411x793mm (72 x 72 DPI)

BMJ Open

Cohort profile: AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353,157 patients in London, United Kingdom

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-058552.R1
Article Type:	Cohort profile
Date Submitted by the Author:	18-Feb-2022
Complete List of Authors:	Wagner, Siegfried; University College London Institute of Ophthalmology, Hughes, Fintan; Duke University Hospital Cortina Borja, Mario; UCL, Institute of Child Health Pontikos, Nikolas; University College London; Moorfields Eye Hospital NHS Foundation Trust Struyven, Robbert; University College London; Moorfields Eye Hospital NHS Foundation Trust Liu, Xiaoxuan; University Hospitals Birmingham NHS Foundation Trust; University of Birmingham Montgomery, Hugh; University College London, Centre for Human Health and Performance Alexander, Daniel; University College London, Department of Computer Science Topol, Eric; Scripps Research Institute Petersen, Steffen; Queen Mary University of London, William Harvey Research Institute; Barts Health NHS Trust, Balaskas, Konstantinos; Moorfields Eye Hospital NHS Foundation Trust; Moorfields Eye Hospital NHS Foundation Trust Hindley, Jack; University College London, Department of Information Governance Petzold, Axel; UCL; Moorfields Eye Hospital NHS Foundation Trust Rahi, Jugnoo; UCL, Great Ormond Street Institute of Child Health; Great Ormond Street Hospital for Children NHS Foundation Trust Denniston, Alastair; Queen Elizabeth Hospital Birmingham, UK; University of Birmingham Keane, Pearse; University College London; Moorfields Eye Hospital NHS Foundation Trust
Primary Subject Heading:	Ophthalmology
Secondary Subject Heading:	Cardiovascular medicine, Epidemiology, Neurology, Ophthalmology
Keywords:	OPHTHALMOLOGY, Medical ophthalmology < OPHTHALMOLOGY, Medical retina < OPHTHALMOLOGY, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Cohort profile: AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353,157 patients in London, United Kingdom

SK Wagner[1, 2], F Hughes[3], M Cortina-Borja[4], N Pontikos[1, 2], RR Struyven[1, 2], X Liu[5, 6], H Montgomery[7], DC Alexander[8], E Topol[9], S Petersen[10, 11], K Balaskas[1, 2], J Hindley[12], A Petzold[1, 2, 13, 14], J Rahi[1, 4, 15, 16], AK Denniston[5, 6], PA Keane[1, 2, *]

*Corresponding author: Pearse A. Keane, MD FRCOphth, NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, 162 City Road, London, United Kingdom. Tel: +44 207 253 3411 Email: pearse.keane1@nhs.net

Word count (excluding title page, abstract, references, figures and tables): 4473 words

[1] Institute of Ophthalmology, University College London, London, UK,

[2] Moorfields Eye Hospital NHS Foundation Trust, London, UK,

[3] Duke University Hospital, Durham, North Carolina, USA,

[4] Great Ormond Street Institute of Child Health, University College London, London, UK,

[5] University of Birmingham, Birmingham, UK,

[6] University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK,

[7] Centre for Human Health and Performance, University College London, London, UK

[8] Centre for Medical Image Computing, Department of Computer Science, University College London, UK

[9] Department of Molecular Medicine, Scripps Research, La Jolla, CA, USA.

[10] William Harvey Research Institute, Queen Mary University of London, London, UK.

[11] St Bartholomew's Hospital, Barts Health National Health Service (NHS) Trust, London, UK..

[12] Department of Information Governance, School of Life and Medical Sciences, University College London, London, UK

[13] The National Hospital for Neurology and Neurosurgery (UCLH), Queen Square Institute of Neurology, University College London, London, UK.

[14] Neuro-ophthalmology Expert Centre, Amsterdam UMC, Departments of Neurology and Ophthalmology, Netherlands.

[15]Great Ormond Street Hospital NHS Foundation Trust, London, United Kingdom.

[16]Ulverscrosft Vision Research Group, University College London, London, UK

Abstract

Purpose: Retinal signatures of systemic disease ('oculomics') are increasingly being revealed through a combination of high-resolution ophthalmic imaging and sophisticated modelling strategies. Progress is currently limited not mainly by technical issues, but by the lack of large labelled datasets, a sine qua non for deep learning. Such data are derived from prospective epidemiological studies, in which retinal imaging is typically unimodal, cross-sectional, of modest number and relate to cohorts, which are not enriched with subpopulations of interest, such as those with systemic disease. We thus linked longitudinal multimodal retinal imaging from routinely-collected National Health Service data with systemic disease data from hospital admissions using a privacy-by-design third-party linkage approach.

Participants: Between January 1st 2008 and April 1st 2018, 353,157 participants aged 40 years or older, who attended Moorfields Eye Hospital NHS Foundation Trust, a tertiary ophthalmic institution incorporating a principal central site, four district hubs and five satellite clinics in and around London, United Kingdom serving a catchment population of approximately six million people.

Findings to date: Among the 353,157 individuals, 186,651 had a total of 1,337,711 Hospital Episode Statistics admitted patient care episodes. Systemic diagnoses recorded at these episodes include 12,022 patients with myocardial infarction, 11,735 with all-cause stroke and 13,363 with all-cause dementia. A total of 6,261,931 retinal images of seven different modalities and across three manufacturers were acquired from 154,830 patients. The majority of retinal images were retinal photographs (n=1,874,175) followed by optical coherence tomography (n=1,567,358).

Future plans: AlzEye combines the world's largest single institution retinal imaging database with nationally collected systemic data to create an exceptional large-scale, enriched cohort that reflects the diversity of the population served. First analyses will address cardiovascular diseases and dementia, with a view to identifying hidden retinal signatures that may lead to earlier detection and risk management of these life-threatening conditions.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- AlzEye is a large retrospective cohort dataset linking ophthalmic data from Moorfields Eye Hospital NHS Foundation Trust in London, United Kingdom with National Health Service hospital admissions data over a ten-year period in 353,157 patients.

- The dataset consists of more than six million routinely collected retinal images of seven different modalities across three vendors deterministically linked to prevalent and incident cardiovascular and neurodegenerative disease.

- Actively informed by ongoing patient and public engagement, the project leverages a privacy-by-design approach using third-party linkage to facilitate access to high-performance computing while mitigating risks to data privacy.

INTRODUCTION

Scientific discovery has increasingly been driven by the availability of large, diverse, high-dimensional datasets providing deeply phenotyping variables in health and disease[1–3]. Advances in healthcare informatics, hardware and statistical techniques have uncovered relationships previously unachievable through traditional methods of study design. Thus, rich and voluminous genome sequencing data has provided insight into disease pathogenesis and therapeutic targets [4,5] while radiomic analysis has supported exploration of relationships between quantitative data extracted from medical imaging and disease [6].

Crucial to health data research has been the establishment of and accessibility to large prospective epidemiological studies, such as the United Kingdom Biobank (UKBB), the Rotterdam study, and the European Prospective Investigation into Cancer and Nutrition (EPIC) study[7–9]. While such studies represent exceptionally powerful enablers for discovery science, they are potentially limited for investigations of specific subpopulations of interest (e.g. those with rare disease or specific sociodemographic groups) and, where they draw on volunteer participants, also prone to selection bias (e.g. over-representation of more healthy subjects). Participants in UKBB are less likely to be obese, smoke, or drink alcohol and accordingly, mortality rates for participants aged 70-74 in UKBB are 46.2% and 55.5% lower for men and women respectively compared to general UK population[10].

Healthcare in England is such that when a patient has a medical event requiring admission, in almost all cases they are admitted under the provisions of the National Health Service (NHS). Routinely collected healthcare administrative data during a patient's admission are subsequently translated into corresponding International Classification of Diseases (ICD) codes by clinical coders, submitted to the Secondary Uses Service and aggregated by NHS Digital into

a unified record-level national repository of Hospital Episode Statistics (HES) data relating admitted patient care (APC). While the original purpose of HES was the monitoring of service activity and negotiation of financial reimbursement, it is increasingly used for epidemiological research[11]. HES data are amenable to research as a sole resource. However, using deterministic linkage where identifiers are matched in a rules-based approach in contrast to probabilistic linkage[12], HES can enrich other datasets as in the case of UKBB[13] or the European Prospective Investigation into Cancer in Norfolk[14].

While these aforementioned studies demonstrate the value of enriching structured datasets through HES linkage, this has not yet been done at scale for routinely collected data of high-dimensionality, such as imaging. Thus, we established AlzEye, a large dataset which links routinely collected retinal images and relevant ophthalmic data from an unselected population attending Moorfields Eye Hospital NHS Foundation Trust (MEH) with nationally collected systemic healthcare outcome data provided through the HES APC database. MEH is a tertiary ophthalmic institution incorporating a principal central site, four district hubs and five satellite clinics in London, United Kingdom (UK), providing care to a sociodemographically diverse population of six million people (9% of the UK population)[15]. The aim of AlzEye is to characterise the association between retinal biomarkers and chronic disorders of ageing, particularly dementia and cardiovascular diseases. In addition to describing the characteristics of the AlzEye cohort, we outline the key governance, technical and ethical factors that need to be addressed to support large institution-led individual-level linkage of routinely collected multi-dimensional data and have enabled us to create an exceptional trans-disciplinary resource to explore the retinal signatures of systemic disease.

COHORT DESCRIPTION

Study population

The AlzEye project is a retrospective cohort study of patients aged 40 years and over who have attended MEH between January 1st 2008 and April 1st 2018. Patients were included if they had attended the glaucoma, retina, neuro-ophthalmology or emergency ophthalmic services and had valid NHS numbers. Those with invalid NHS numbers, dates of birth or who had previously opted out of their health data being used for purposes of research (described in the NHS as a 'Type 2 opt-out') were excluded. Ethnicity group was self-reported by the patient as i) Asian or Asian British, ii) Black or Black British, iii) Mixed, iv) Other Ethnic Group, v) White, or vi) Unknown. Socio-economic status was categorised using the Index of multiple deprivation (IMD) decile, which was estimated by permuting the IMD 2015 rank from the patient's postcode through Lower Super Output Areas followed by aggregation into deciles[16]. Mortality data were derived from the MEH database, which is updated on a two-weekly basis using reports extracted from the NHS National Spine, and is completed on an individual basis by the MEH Data Quality team to ensure accuracy. Data are completed on any patients who have ever attended MEH. Mortality data up to the end of the study period, 1st April 2018, were included.

Approvals and process

The following key steps in the governance processes were required to provide the necessary ongoing assurance within the research ethics framework of the NHS and the legal framework of the UK. In order to support other researchers wishing to establish similar linked cohorts, we provide an explanation of each stage which outlines the principle which that stage addresses, the UK framework that meets that principle, and finally any study-specific considerations that we undertook to not only meet but exceed those requirements (Figure 1).

- Funding

It was necessary to secure funding to deliver the study, and to provide assurance to the sponsor and others that the study would be completed and that the integrity of the study would not be compromised by inadequate resources. In AlzEye, the study was funded through a small grant awarded by Fight for Sight and Alzheimer's Research UK in October 2017 covering the costs of data storage and linkage fees. The funders had no role in the conception, design or analysis of the study.

- Sponsorship

It was necessary to secure a sponsor for the study that would take on 'overall responsibility for proportionate, effective arrangements being in place to set up, run and report on a research project'. In AlzEye, we sought sponsorship from the relevant NHS Trust (MEH) at which all patients had been seen. Sponsorship confirmation was only sought following internal consultation between data protection, information governance, information security and information technology (IT) teams at both MEH and UCL. MEH acted as the Sponsor, with UCL acting as the trusted third party linking retinal images and HES data and providing

computational facilities for data analysis. The study and data governance were approved on 24/05/2018 (internal reference: KEAP1004).

- Health Research Authority (HRA) Approval

For most research studies in England and Wales, including those limited to working with data for specific projects, Health Research Authority (HRA) approval is required. HRA approvals involve the assessment of governance and legal compliance of a research study with an independent ethics review and opinion from the Research Ethics Committee (REC). Depending on other study characteristics (e.g.gene therapy), additional applications may be required to inform HRA approval. In England, limited access to confidential patient information without consent may be granted under the provisions of Section 251 of the National Health Service Act 2006, permitting temporary lifting of the common law duty of confidentiality around confidential patient information 'in the public interest' or 'in the interests of improving patient care'[17]. Obtaining section 251 support requires application to the Confidential Advisory Group (CAG), an independent body providing expert advice to the HRA for research applications and NHS Digital for data dissemination. Applications were accordingly made to the Research Ethics Committee (18/LO/1163, approved 01/08/2018) and the CAG for Section 251 support (18/CAG/0111, approved 13/09/2018). The National Health Service Health Research Authority gave final approval on 13/09/2018. Approvals thus far granted the legal basis for submitting an application to the Data Access Request Service (DARS) of NHS Digital[18].

- NHS Digital and the Data Access Request Service (DARS)

NHS Digital oversees the Data Access Request Service (DARS), which administers and provides, upon application, multiple England-wide datasets from disease-specific audits (e.g. National Diabetes Audit Core) to general admissions in secondary care (e.g. HES). Applications to DARS require that the organisation have, at a minimum, the following:

i) Data Sharing Framework Contract for Data Controllers

ii) Compliance with minimum-security standards for Data Processors and Data Storage locations

iii) Adequate information security certification (e.g. ISO27001)

iv) A legal basis for data access (e.g. Section 251)

Applications are then reviewed with an assigned case officer, who will liaise with the applicant on project-specific items. For AlzEye, dialogue between the applicant and NHS Digital data production team revolved around confirmation of data fields and datasets (HES) and the pseudonymisation embedded within the linkage strategy.

- Independent Group Advising on the Release of Data (IGARD)

Following internal NHS Digital review and prior to data release, DARS applications are scrutinised by the Independent Group Advising on the Release of Data (IGARD) in line with Section 263(2) of the Health and Social Care Act 2012, the Code of Practice on confidential information. IGARD is an independent panel with a broad range of expertise, from legal to information governance to epidemiology. Support for AlzEye was given by IGARD in January and August 2019 citing that 'aspects of the application could be used as an exemplar by NHS Digital to help other researchers with their applications to the Data Access Request Service (DARS)'[19].

- Data sharing agreements (DSA)

Prior to data receipt, data sharing agreements (DSA) must be signed between NHS Digital and the Data Controller and are overseen by their respective legal departments. AlzEye required an additional DSA between MEH and UCL for the transfer of ophthalmic imaging and clinical data between institutions outlining the purpose and legal basis for sharing.

- Data processing and transfer

The dataset was finalised upon completion of engineering work parsing manufacturer-specific file formats to non-proprietary data structures amenable to image analysis with appropriate deidentification. A secure cloud-based informatics pipeline was used for transfer of images to UCL from MEH, the establishment of which was delayed by the COVID-19 pandemic. Imaging data was stored (with backup) across dedicated network-attached storage device within the UCL School of Life and Medical Sciences (SLMS) and only accessible to members of the AlzEye research team. All data entities were listed within the UCL SLMS Information Asset Register.

Patient and public involvement and engagement

Patient and public involvement and engagement (PPIE) support was provided by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at Moorfields Eye Hospital and UCL Institute of Ophthalmology and has been embedded throughout this project from priority setting to plans for dissemination. Feedback has been sought through public engagement events, survey of eye service users and reports within the media. Patients and the public actively contributed to identifying the priority setting of dementia and the acceptability of using routinely acquired eye scans for research purposes and without consent. In addition, two members of the public will sit on the AlzEye working group to contribute to results interpretation and co-authoring and dissemination of research outputs. The members will be supported in selecting the results they find relevant and presenting them to wider patient communities.

Ophthalmic health variables

Patient-level ophthalmic variables were extracted from the MEH data warehouse, which aggregates information from the patient administration system (PAS), electronic health record and imaging database, all linked through a unique MEH hospital identification number. Sociodemographic data, including date of birth, sex, ethnicity and postcode as well as patients' clinic appointments and operation dates are housed within PAS. Surgical procedures were recorded in the electronic health record (EHR) at MEH from September 4th, 2012. Operation details, including procedure name, laterality and indication for surgery are contained within the MEH EHR and uploaded to the MEH data warehouse. A patient undergoing the most common operation in the UK, cataract extraction, would therefore have an entry for the typical procedure, (*phacoemulsification and intraocular lens implant*) operated eye (*right or left*) and indication (*cataract*).

Colour retinal photography (Figure 2A) and optical coherence tomography (OCT, Figure 2B) images, which represent the majority of retinal images within the database, have been processed through segmentation and feature extraction software. The Vascular Assessment and Measurement Platform for Images of the Retina (VAMPIRE) system provides fully automated segmentation and extraction of retinal vascular indices[20,21]. OCT scans are segmented and retinal sublayer thicknesses computed using the Topcon Advanced Biomedical Imaging Laboratory (TABIL) software[22].

For the purposes of this report, four common ophthalmic diseases were described - cataract, glaucoma, neovascular age-related macular degeneration (AMD) and proliferative diabetic retinopathy (PDR).

Cataract was defined as any operation code denoting phacoemulsification surgery and the indication of cataract. For the purposes of this report, only first eye cataract surgery was included.

Glaucoma was defined as any patient attending the glaucoma clinic three or more times with ongoing follow up from January 1st 2010. The first two years of the study period were excluded as this may have incorporated patients with previous diagnoses of glaucoma where the maximum follow up interval can approach two years; in contrast any patient being seen after 2 years since study inception with no previous visit within that 2 year period can be assumed to have/carry a new diagnosis of glaucoma.

Diabetic eye disease represents a special case due to audit procedures mandated by the NHS Diabetic Eye Screening Programme. Coding of eye disease secondary to diabetes mellitus is rigorously validated by a dedicated team within MEH according to the NHS Diabetic Eye Screening Programme criteria[23], at hospital appointment from September 12, 2013 onwards. Dates for onset of proliferative diabetic retinopathy dates were recorded as the first appointment for each patient where this diagnosis was first made.

AMD can be categorised into two major types - dry and neovascular ("wet"). Given dry AMD is slowly progressive and has no active hospital intervention currently available, it is MEH standard practice for patients to be discharged with lifestyle and monitoring advice (self-monitoring and standard optometric review). In contrast, neovascular AMD requires treatment through intravitreal anti-VEGF injections, and therefore remains under active follow-up. The diagnostic codes for neovascular AMD were based on extensive previous work in which all patients with neovascular AMD at MEH were manually validated up to 2018[24,25].

Systemic health variables

Systemic health data were derived from HES APC data, with a focus on cardiovascular disease and all-cause dementia. Diagnostic codes in HES APC are reported in line with the 10th revision of ICD[26]. In line with previous reports, myocardial infarction was defined as code I21 or I22[27–29]. Similarly, stroke was defined using stroke definitions from UK Biobank[30]. Dementia was defined as ICD codes E512 (Wernicke's encephalopathy), F00 (Dementia in Alzheimer disease), F01 (Vascular dementia), F02 (Dementia in other diseases classified elsewhere), F03 (Unspecified dementia), F10.6 (Mental and behavioural disorders due to psychoactive substance use, Amnesic syndrome) F10.7 (Mental and behavioural disorders due to psychoactive substance use, Residual and late-onset psychotic disorder) , G30 (Alzheimer disease), or G31.0 (Other degenerative diseases of nervous system, not elsewhere classified), derived from previous work evaluating the agreement between HES admitted patient care data and primary care data, through general practitioner surveys and the Clinical Practice Research Datalink (CPRD)[31].

Data Linkage and Transfer

The linkage strategy was designed through collaboration between experts in information governance, information technology, computer scientists and clinicians based at MEH, University College London (UCL) and NHS Digital (Figure 3). A third-party linkage approach was used for two main reasons. Firstly, it enhanced privacy-preservation as the data originator, MEH, never received HES admissions data and the third party, UCL, did not receive personally identifiable information. Secondly, it enabled the linked dataset to be accessible within a site with sufficient high-performance computing capability to undertake the proposed analyses, a function significantly beyond almost all NHS facilities. Patient link identifiers consisting of a

unique NHS identification number, sex and date of birth originating from MEH were transferred to NHS Digital in conjunction with a unique study ID generated using a cryptographic hash function (random pseudonymisation). Ophthalmic covariates, mortality data, and patient sociodemographics with study ID were transferred to UCL. Ophthalmic imaging data pertaining to the patients within the study were extracted and de-identified during conversion from their proprietary format to Digital Imaging and Communications in Medicine (DICOM) format before transfer to UCL. Following linkage with HES, NHS Digital transferred HES data to the UCL Data Safe Haven, a "walled garden" trusted research environment certified and externally audited to ISO27001 information security standards[32,33].

Statistical analysis

Imaging-based studies within the AlzEye study are generally planned to take the form of nested case-control studies. To improve efficiency, controls may be matched with cases, using conditional logistic regression for statistical modelling of binary outcomes and survival analysis for time-to-event data (e.g. Cox Proportional hazard modelling)[34]. In cases where the competing risk of death is prominent, subdistribution hazard ratios with 95% confidence intervals will be estimated as a sensitivity analysis [35]. Alternative high-dimensional modelling approaches, such as vision transformers, will also be explored. Prior to receipt of HES data from NHS Digital, sample size calculations were undertaken. Specifically, we evaluated the association between OCT-derived peripapillary retinal nerve fibre layer and macular ganglion cell-inner plexiform layer thicknesses and dementia. Given an odds ratio of 1.4 with an alpha of 5% and a power of 90% on a 1:1 matched study design, a total sample size of 2106 is required[36].

Figures for this report were designed in R version 4.1.0 (R Core Team, 2021. R Foundation for Statistical Computing, Vienna, Austria).

FINDINGS TO DATE

Extraction of unique patients attending MEH outpatient clinics between January 1st 2008 and April 1st 2018 generated a cohort of 353,157 unique patients. A breakdown of sociodemographic details by category of the cohort are provided in Table 1. Of the cohort, 190,494 were female (53.9%) and the mean age was 68.4 (+/- 13.9) years. Of the 353,157 patients, 186,651 had a total of 1,337,711 HES episodes in the study period. NHS Digital performs a hierarchical stepwise linkage approach providing a 'Match Rank' for each HES episode[37]. Among the 1,337,711 HES episodes matched, Match Rank was two for 1,337,482 episodes (exact NHS number, exact date of birth and exact sex linked), four for 46 episodes (exact NHS number, exact sex and partial date of birth), and eight for 183 episodes (exact NHS number).

An illustration of the major common ophthalmic diseases within the cohort is shown in a CONSORT-style diagram in Figure 4. Following the case definition, and exclusion of invalid dates, a total of 59,102 patients had first eye cataract surgery, 31,060 glaucoma, 7214 neovascular AMD and 2494 PDR.

	Characteristic	N (%)

	All	353,157
Sex	**Female**	190,494 (53.9)
	Male	162,663 (46.1)
Age group*	**40-49 years**	35,262 (10.0)
	50-59 years	66,101 (18.7)
	60-69 years	79,018 (22.4)
	70-79 years	84,942 (24.1)
	80+ years	87,834 (24.9)
Ethnicity	**Black**	31,614 (9.0)
	White	135,743 (38.4)
	South Asian	48,119 (13.6)
	Other/Unknown	137,681 (39.0)
Index of multiple deprivation decile	**1 (most deprived)**	18,194 (5.2)
	2	50,443 (14.3)
	3	50,869 (14.4)
	4	42,603 (12.1)
	5	38,964 (11.0)
	6	36,906 (10.5)
	7	31,317 (8.9)
	8	28,180 (8.0)
	9	29,906 (8.5)
	10 (least deprived)	24,610 (7.0)
	Unknown	1165 (0.3)

Table 1: Baseline sociodemographic characteristics of the AlzEye cohort. Data are shown as n(%).

*Age is taken as that of April 1st, 2018.

Among the 187,811 patients with recorded HES episodes, 12,022 patients had episodes with coded myocardial infarction, 11,735 patients with all-cause stroke and 13,363 with dementia. Within the dementia group, 4487 patients had codes that were specific for Alzheimer's dementia and 3381 for vascular dementia (Table 2).

Group	Disease	ICD code(s)	Number of patients
Cardiovascular	Acute coronary syndrome	I21, I22	12,022
	Heart failure	I50	24,034
	Atrial Fibrillation	I48	32,848
	Hypertension	I10, I15	151,937
	Subarachnoid haemorrhage	I60	642
	Intracerebral hemorrhage	I61	1865
	Ischaemic stroke	I63-I64	9996
	All stroke	I60, I61, I63, I64	11,735
Neurodegenerative	Alzheimer's disease	F00, G30	4487
	Vascular dementia	F01	3381
	Parkinson's disease	G20	3211
	All-cause dementia	E12, F00, F01, F02, F03, F106, F107, G30, G310	13,363
Other	Diabetes mellitus (Type 1 and 2)	E10, E11	71,570

Table 2: Number of patients by selected examples of specified 10th revision of International Classification of Diseases (ICD) codes relating to diabetes mellitus, cardiovascular and neurodegenerative diseases.

Imaging

During the study period, a total of 6,261,931 images were acquired from 154,830 patients. The two leading image modalities were colour retinal photographs (n=1,874,175) and OCT (n=1,567,358). The distribution of imaging modalities across the three vendors used for retinal imaging at MEH - Topcon (Topcon corp, Tokyo, Japan), Heidelberg (Heidelberg Engineering, Heidelberg, Germany), and Optos (Optos, Dunfermline, UK) - are shown in Figure 5 and Table

2. Most images were acquired on the Topcon system (n=5,553,826, 88.7%). Number of images by year is shown in Figure 6. During the study period, annual imaging acquisition increased from 229,868 scans in 2008 to 1,021,904 in 2017. For 2018, collection stopped on April 1st precluding a complete annual figure. Example images of the major ophthalmic and systemic disease outcomes are shown in Figure 7.

Vendor	Modality	Number of images	Number of patients
Topcon	Angiography	1,128,723	21,225
	Autofluorescence	11,761	2078
	Colour photography	1,874,175	139,307
	Red-free	1,146,854	122,453
	OCT	1,391,826	138,911
	Other	487	48
Heidelberg	Angiography	89,264	4061
	Autofluorescence	94,533	16,863
	Infrared	192,634	21,676
	OCT	175,532	21,191
	Other	19,781	2439
Optos	Angiography	77,813	2215
	Autofluorescence	18,590	5666
	Pseudocolour photography	39,958	6887

Table 3: Retinal imaging within the AlzEye dataset by vendor and imaging modality. OCT: Optical coherence tomography. Angiography refers to dye-based techniques (fluorescein and indocyanine green).

STRENGTHS AND LIMITATIONS

To our knowledge, we have created the world's largest retinal imaging research dataset available presently, linking secondary healthcare ophthalmic data from 353,157 patients seen over a ten-year period with information on general health and key systemic diseases, as captured through admissions to any hospital within the NHS of England. This comprises 6,261,931 images, obtained using seven different modalities from three different manufacturers, in 154,830 patients. The current large-scale UK cohort, UKBB, provides useful context for AlzEye. Cross-sectional data are available in UKBB with two retinal imaging modalities (colour retinal photography and OCT) obtained using technology from one manufacturer (Topcon), and at a single time point in 67,321 people. Notwithstanding the recognised limitations (see 'Limitations' section) of real-world datasets and the coding within the HES database, AlzEye provides some distinct advantages beyond purely scale. Imaging data are longitudinal, highly multimodal, and pertain to an ethnically and socioeconomically diverse cohort representative of the adult population with eye disease. Moreover, AlzEye has demonstrated relatively low cost. The study is funded through a charity small grant award and NIHR Biomedical Research Centre support amounting to £20,000.

Comparison with other resources

UKBB is the major comparator for AlzEye, being the largest of the prospective epidemiological cohort datasets which provide cross-sectional retinal imaging in association with systemic disease variables[38]. One of the limitations of UKBB is that, unlike AlzEye, it provides minimal longitudinal retinal images. Another prospective cohort study, the Rotterdam study, does collect longitudinal retinal imaging data from approximately 15,000 participants, of which 5065 participants were eligible for OCT scanning in 2017[39]. The Rotterdam Study has uncovered several landmark findings, particularly in regards to causal determinants, but its cohort remains

relatively small in comparison to UKBB and AlzEye with the majority of participants recruited from one district within Rotterdam, Netherlands[7,40]. The Singapore Epidemiology of Eye Disease (SEED), is one longitudinal multimodal retinal imaging initiative which is underway, in which 10,033 participants of Chinese, Indian and Malay ethnicity have been recruited to undergo six-yearly retinal imaging[41]. A recent review of ophthalmic imaging datasets did not reveal any additional relevant publicly available datasets that included linked systemic health data[42]. Additionally, our own review of the literature has not identified any examples of large-scale linked real-world datasets (i.e. including those with restricted access) which include linked systemic health data. The scarcity of such resources suggests that the construction of such datasets is challenging to undertake, presumably due to factors such as cost, required duration and delayed output, retention of participants, and concerns over technological redundancy. The AlzEye approach is an important alternative model in this context.

Potential research impact from the novel AlzEye cohort

Several epidemiological opportunities arise with AlzEye. Firstly, it provides a real-world snapshot of ophthalmic secondary care use, representing approximately 1.2% of the UK population aged 40 and above (27,858,459 in 2011) [43]. This is a powerful tool for informing public health and policymaking in eye services, and is exceptional in characterising the potential impact that may arise from the intersect between disabling diseases such as stroke and PDR.

Secondly, it allows the identification and exploration of relationships between newly diagnosed ophthalmic disease (or newly referred to hospital eye services) and emerging systemic events and accruing multimorbidity. Patients tend to respond early to issues with their sight and an understanding of how an ophthalmic presentation is linked to an increased likelihood of serious systemic disease may provide an opportunity for earlier intervention in those diseases[44].

Thirdly, nested case-control studies evaluating retinal-based oculomic biomarkers in those with systemic diseases (e.g. dementia) can provide insight into their value in either static or dynamic risk prediction. Newer modelling approaches have highlighted the potential utility of the retina in screening for and risk stratification of cardiovascular, neurodegenerative, renal, hepatic and haematological diseases[45–50];[51].

Finally, by its magnitude and wealth of high-quality labels, both ophthalmic and systemic, AlzEye provides a powerful catalyst for high-dimensional model development, echoing that of Imagenet, a database currently exceeding 14 million images, which propelled deep learning and computer vision research forward a decade ago[52].

Lessons learned from the AlzEye approach

AlzEye highlights an opportunity for maximising the value of routinely-collected data to support research for patient benefit. However, there are a number of governance and technical challenges when undertaking large scale investigator-led data linkage[53]. In AlzEye, early dialogue between experts in information governance, information technology and data protection at both institutional parties (MEH and UCL) as well as the data production team at NHS Digital established a privacy-by-design linkage approach, which enhanced privacy preservation while maintaining computational feasibility[30,54]. At its worst, an intrusion of the identifiable data during the development of AlzEye would have informed the violator that a given individual had visited MEH at some point between 2008 and 2018. Due to the novel approach of AlzEye within our centre, the greatest governance hurdle was securing study sponsorship, a process which took nearly eight months. Once approved, permissions from the bodies of the Health Research Authority, and overall approval were given within eight weeks. Linking with high-dimensional imaging data also posed several technical obstacles. As highlighted recently

by the American Academy of Ophthalmology (AAO), ophthalmic imaging technologies suffer from limited interoperability and low compliance to standardised formats, such as DICOM [55]. A key undertaking within AlzEye was thus the secure and robust but efficient fully-automated processing of raw ophthalmic imaging data from its proprietary file format with associated metadata to standard DICOM form with the identifiers stripped. Fortunately, while this operation requires significant technical and engineering input, most medical imaging modalities already benefit from standardisation among vendors obviating this step for other researchers seeking to emulate our approach. Finally, a key objective of AlzEye is the development of clinical prediction models using deep learning approaches, which require significant computing capacity. Provisions for graphics processing units (GPU) housed within UCL enable this step however others may consider recent guidance on the safeguards required for locating health data within cloud environments and the implications this brings for accessing virtual GPUs[56].

Limitations of the AlzEye cohort

Despite the opportunities afforded by AlzEye, there are several limitations to this kind of approach and potential sources of bias. Firstly, caution must be paid to the validity of HES diagnostic coding[57]. Although previous validation studies have concluded that discharge coding within HES is sufficiently robust for research purposes[31,58], sizable proportions of cases may be missed when using individual sources[59]. For example, recent work linking the electronic health records of 54.4 million people in England showed that HES captured 80.5% and 65% of myocardial infarctions and stroke/transient ischaemic attacks respectively when compared to linkage additionally incorporating Death Registry and primary care records[60]. One mitigation strategy for this source of bias for real world data is therefore linking to multiple sources. In terms of selection bias, as a hospital-attending cohort, the individuals within the AlzEye cohort are likely to have greater medical comorbidity than the general population, limiting the external validity of any findings. In addition, by the very nature of the dataset,

patients within the AlzEye cohort will have definite or suspected ophthalmic disease, particularly among those with repeated retinal imaging. The risk of under-recording of potentially important variables such as smoking may also lead to residual confounding.

The enrichment of multimodal health data acquired as part of a patient's routine clinical care with nationally-held databases provides a powerful foundation for discovery science and epidemiological research. We highlight key considerations and challenges for those seeking to link high-dimensional data sources, from high-resolution imaging to waveform data, with locally-held specialist data. Additionally, we provide the cohort profile for AlzEye, a powerful platform for oculomic discovery, specifically evaluating the association between retinal morphology, and both cardiovascular diseases and dementia. Beyond discovery, the AlzEye cohort is anticipated to become an important resource for the development and validation of deep learning-based clinical prediction models that may enable earlier intervention for patients at risk of these life-threatening conditions.

LEGENDS

Figure 1: Schematic of the key milestones, prerequisites and approvals with their corresponding achievement dates for the AlzEye dataset. REC: Research Ethics Committee, CAG: Confidential Advisory Group, HRA: Health Research Authority, NHS: National Health Service, IGARD: Independent Group Advising on the Release of Data, DSA: Data Sharing Agreement.

Figure 2: Composite figure showing the major retinal imaging modalities within AlzEye. A: Colour fundus photograph, B: Red-free photograph, C: Fundus autofluorescence (widefield), D: Pseudocolour photography (widefield) and E: Optical coherence tomography of the central macula illustrating segmentation of the individual sublayers. Consensus nomenclature for the retinal sublayers is indicated. NFL= nerve fibre layer, GCL = ganglion cell layer, IPL: inner plexiform layer, INL: inner nuclear layer, OPL: outer plexiform layer, ONL = outer nuclear layer, MZ = myoid zone, RPE-IZ = RPE-interdigitation zone, BM = Bruch's membrane.

Figure 3: Linkage approach of AlzEye. Moorfields Eye Hospital NHS Foundation Trust securely transfers a spreadsheet of identifiers with a study ID to NHS Digital and separately transfers the study ID with ophthalmic data, including diagnoses and retinal images, to University College London. NHS Digital links the identifiers with the Hospital Episode Statistics (HES) database and returns the admissions data with the study ID (and no identifiable data) to UCL. UCL links the ophthalmic data from Moorfields Eye Hospital with HES data from NHS Digital using the study ID.

Figure 4: CONSORT style flow chart illustrating the distribution of cataract, glaucoma, neovascular age-related macular degeneration (AMD) and proliferative diabetic retinopathy (PDR) within the AlzEye dataset.

Figure 5: Parallel sets diagram illustrating the imaging modality across vendors within AlzEye. The majority of images were acquired on the Topcon system and the most frequent modalities were colour photography and optical coherence tomography. Designed using the `networkD3` package.

Figure 6: Stacked bar chart of the annual number of images acquired during the study period for the three leading device vendors at Moorfields Eye Hospital. Data for 2018 represents 3 months only prior to the study end-date.

Figure 7: Example colour retinal photographs of ophthalmic and systemic diseases within AlzEye. A: age-related macular degeneration, B: cataract, C: glaucoma, D: proliferative diabetic retinopathy E: prevalent Alzheimer's disease, F: incident ischaemic stroke, G: incident myocardial infarction, H: prevalent vascular dementia.

ACKNOWLEDGEMENTS

We thank Karen Bonstein and Andi Skilton for support with patient and public involvement and engagement, Menachem Katz, Ben Ward, Ross Green, Maxim Daniline and Simon St John-Green for information technology support, Anthony Peacock for advice on use of Trusted Research Environments and Antonio de la Plaza Larrea and Richard Macmillan for legal guidance on data sharing agreements. We also are grateful to Anthony Khawaja and Cathie Sudlow for feedback on study design.

CONTRIBUTORS

SW, AD and PK wrote the first draft of the manuscript, which was critically revised by FH, MCB, RS, NP, XL, HM, DA, ET, SP, KB, JH, AP, and JR. SW, FH, NP, HM, JH, AP, JR, AD and PK were involved in the original design of the study. Computer science expertise was through RS, NP and DA, information governance through JH. MCB, AP, JR and AD provided statistical and epidemiological guidance. All authors have approved the final version of this manuscript.

FUNDING STATEMENT

The study was funded through a small grant awarded by Fight for Sight (grant reference: 24AZ171). Infrastructural support was through the NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology. The funders had no role in the conception, design or analysis of the study.

COLLABORATION

National and international collaborations are welcomed though restrictions on access to the cohort mean that only the AlzEye researchers can directly analyse individual-level systemic health data. Interested researchers should contact s.wagner@ucl.ac.uk.

DATA AVAILABILITY STATEMENT

No additional data available. The data are subject to the contractual restrictions of the DSA between NHS Digital, Moorfields Eye Hospital and University College London and are therefore not available for access beyond the AlzEye research team.

COMPETING INTERESTS STATEMENT

SW is funded through a Medical Research Council Clinical Research Training Fellowship (MR/TR000953/1).

FH: None.

MCB: None.

NP is funded by a Moorfields Eye Charity Career Development Award (R190031A).

RS: None.

XL: None.

HM is supported by the National Institute of Health Research's Comprehensive Biomedical Research Centre (BRC) at University College London Hospitals (UCLH).

DA: None

ET: None

SP receives support from the National Institute for Health Research Biomedical Research Centre at Barts.

KB has received speaker fees from Novartis, Bayer, Alimera, Allergan, Roche, and Heidelberg; meeting or travel fees from Novartis and Bayer; compensation for being on an advisory board from Novartis and Bayer; consulting fees from Novartis and Roche; and research support from Apellis, Novartis, and Bayer.

JH: None

AP receives financial support from the National Institute for Health Research Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. AP is part of the steering committee of the ANGI network which is sponsored by ZEISS, steering committee of the OCTiMS study which is sponsored by Novartis, and reports speaker fees from Heidelberg Engineering.

JR receives support from the National Institute for Health Research as a Senior Investigator and via the National Institute for Health Research Biomedical Research Centres at Moorfields Eye Hospital and Great Ormond Street Hospital.

AD is Director of INSIGHT, the HDRUK Health Data Research Hub for Eye Health.

PK is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1); receives research support from Apellis; is a consultant for DeepMind, Roche, Novartis, Apellis, and BitFount; is an equity owner in Big Picture Medical; and has received speaker fees from Heidelberg Engineering, Topcon, Allergan, Roche, and Bayer; meeting or travel fees from Novartis and Bayer; and compensation for being on an advisory board from Novartis and Bayer.

ETHICAL APPROVAL STATEMENT

This study was approved by the Research Ethics Committee (18/LO/1163, approved 01/08/2018) and the CAG for Section 251 support (18/CAG/0111, approved 13/09/2018). The National Health Service Health Research Authority gave final approval on 13/09/2018.

REFERENCES

1 Munevar S. Unlocking Big Data for better health. *Nat Biotechnol* 2017;**35**:684–6.

2 Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* 2020;**26**:29–38.

3 Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;**375**:1216–9.

4 Bhardwaj R, Sethi A, Nambiar R. Big data in genomics: An overview. In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE 2014. doi:10.1109/bigdata.2014.7004392

5 He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. *Int J Mol Sci* 2017;**18**. doi:10.3390/ijms18020412

6 Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;**278**:563–77.

7 Hofman A, Breteler MMB, van Duijn CM, *et al.* The Rotterdam Study: objectives and design update. *Eur J Epidemiol* 2007;**22**:819–29.

8 Riboli E, Hunt KJ, Slimani N, *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;**5**:1113–24.

9 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9.

10 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.

11 Chaudhry Z, Mannan F, Gibson-White A, *et al.* Research Outputs of England's Hospital Episode Statistics (HES) Database: Bibliometric Analysis. Journal of Innovation in Health Informatics. 2017;**24**:329. doi:10.14236/jhi.v24i4.949

12 Zhu Y, Matsuyama Y, Ohashi Y, *et al.* When to conduct probabilistic linkage vs.

deterministic linkage? A simulation study. *J Biomed Inform* 2015;**56**:80–6.

13 Hospital information boosts UK Biobank resource. https://www.ukbiobank.ac.uk/2013/09/20000-participants-return-for-a-repeat-assessment/ (accessed 13 Oct 2020).

14 Luben R, Hayat S, Khawaja A, *et al.* Residential area deprivation and risk of subsequent hospital admission in a British population: the EPIC-Norfolk cohort. BMJ Open. 2019;**9**:e031251. doi:10.1136/bmjopen-2019-031251

15 Large P. Annual mid-year population estimates, UK - Office for National Statistics. 2014.https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/2014-06-26 (accessed 23 Jul 2021).

16 Ministry of Housing, Communities & Local Government. English indices of deprivation 2015. 2015.https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 (accessed 18 Sep 2020).

17 Great Britain. *National Health Service Act 2006*. The Stationery Office 2006.

18 Data Access Request Service (DARS). https://digital.nhs.uk/services/data-access-request-service-dars (accessed 7 Jul 2021).

19 Independent Group Advising on the Release of Data. https://webarchive.nationalarchives.gov.uk/20200706184649/https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data (accessed 7 Jul 2021).

20 Perez-Rovira A, MacGillivray T, Trucco E, *et al.* VAMPIRE: Vessel assessment and measurement platform for images of the REtina. *Conf Proc IEEE Eng Med Biol Soc* 2011;**2011**:3391–4.

21 Liew G, Wang JJ, Cheung N, *et al.* The retinal vasculature as a fractal: methodology, reliability, and relationship to blood pressure. *Ophthalmology* 2008;**115**:1951–6.

22 Keane PA, Grossi CM, Foster PJ, *et al.* Optical Coherence Tomography in the UK Biobank Study - Rapid Automated Analysis of Retinal Thickness for Large Population-Based Studies. *PLoS One* 2016;**11**:e0164095.

23 Peate I. The NHS diabetic eye screening programme. British Journal of Healthcare Assistants. 2019;**13**:596–9. doi:10.12968/bjha.2019.13.12.596

24 Fasler K, Fu DJ, Moraes G, *et al.* Moorfields AMD database report 2: fellow eye involvement with neovascular age-related macular degeneration. *Br J Ophthalmol* 2020;**104**:684–90.

25 Fasler K, Moraes G, Wagner S, *et al.* One- and two-year visual outcomes from the Moorfields age-related macular degeneration database: a retrospective cohort study and an open science resource. *BMJ Open* 2019;**9**:e027441.

26 ICD-10 Version:2010. https://icd.who.int/browse10/2010/en (accessed 22 Jul 2021).

27 Asaria P, Elliott P, Douglass M, *et al.* Acute myocardial infarction hospital admissions and deaths in England: a national follow-back and follow-forward record-linkage study. The Lancet Public Health. 2017;**2**:e191–201. doi:10.1016/s2468-2667(17)30032-4

28 Metcalfe A, Neudam A, Forde S, *et al.* Case Definitions for Acute Myocardial Infarction in Administrative Databases and Their Impact on In-Hospital Mortality Rates. Health Services Research. 2013;**48**:290–318. doi:10.1111/j.1475-6773.2012.01440.x

29 McCormick N, Lacaille D, Bhole V, *et al.* Validity of myocardial infarction diagnoses in administrative databases: a systematic review. *PLoS One* 2014;**9**:e92286.

30 [No title]. https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/alg_outcome_stroke.pdf (accessed 23 Aug 2021).

31 Brown A, Kirichek O, Balkwill A, *et al.* Comparison of dementia recorded in routinely collected hospital admission data in England with dementia recorded in primary care. *Emerg Themes Epidemiol* 2016;**13**:11.

32 Lea NC, Nicholls J, Dobbs C, *et al.* Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. JMIR Medical Informatics. 2016;**4**:e22. doi:10.2196/medinform.5571

33 Certificate Client Directory search results. https://www.bsigroup.com/en-GB/our-services/certification/certificate-and-client-directory/search-results/?searchkey=licence%3dIS%2b612909%26company%3duniversity%2bcollege%2bLondon&licencenumber=IS%20612909 (accessed 26 Jul 2021).

34 Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;**34**:187–202. doi:10.1111/j.2517-6161.1972.tb00899.x

35 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;**94**:496–509. doi:10.1080/01621459.1999.10474144

36 Mutlu U, Colijn JM, Ikram MA, *et al.* Association of Retinal Neurodegeneration on Optical Coherence Tomography With Dementia: A Population-Based Study. *JAMA Neurol* 2018;**75**:1256–63.

37 Harper G. Linkage of Maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005–2014: Quality assurance of linkage of routine data for singleton and multiple births. BMJ Open. 2018;**8**:e017898. doi:10.1136/bmjopen-2017-017898

38 Chua SYL, Thomas D, Allen N, *et al.* Cohort profile: design and methods in the eye and vision consortium of UK Biobank. *BMJ Open* 2019;**9**:e025077.

39 Mutlu U, Bonnemaijer PWM, Ikram MA, *et al.* Retinal neurodegeneration and brain MRI markers: the Rotterdam Study. *Neurobiol Aging* 2017;**60**:183–91.

40 Ikram MA, Brusselle GGO, Murad SD, *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* 2017;**32**:807–50.

41 Majithia S, Tham Y-C, Chee M-L, *et al.* Cohort Profile: The Singapore Epidemiology of Eye

Diseases study (SEED). *Int J Epidemiol* 2021;**50**:41–52.

42 Khan SM, Liu X, Nath S, *et al.* A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021;**3**:e51–66.

43 Age groups. 2018.https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest#:~:text=by%20ethnicity%20Summary-,The%20data%20shows%20that%3A,aged%2060%20years%20and%20over (accessed 15 Jul 2021).

44 Enoch J, McDonald L, Jones L, *et al.* Evaluating Whether Sight Is the Most Valued Sense. *JAMA Ophthalmol* 2019;**137**:1317–20.

45 Sabanayagam C, Xu D, Ting DSW, *et al.* A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *Lancet Digit Health* 2020;**2**:e295–302.

46 Mitani A, Huang A, Venugopalan S, *et al.* Detection of anaemia from retinal fundus images via deep learning. *Nat Biomed Eng* 2020;**4**:18–27.

47 Poplin R, Varadarajan AV, Blumer K, *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;**2**:158–64.

48 Wisely CE, Wang D, Henao R, *et al.* Convolutional neural network to identify symptomatic Alzheimer's disease using multimodal retinal imaging. *Br J Ophthalmol* Published Online First: 26 November 2020. doi:10.1136/bjophthalmol-2020-317659

49 Cheung CY, Xu D, Cheng C-Y, *et al.* A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nat Biomed Eng* 2021;**5**:498–508.

50 Xiao W, Huang X, Wang JH, *et al.* Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. *Lancet Digit Health* 2021;**3**:e88–97.

51 Wagner SK, Fu DJ, Faes L, *et al.* Insights into Systemic Disease through Retinal Imaging-Based Oculomics. *Transl Vis Sci Technol* 2020;**9**:6.

52 Deng J, Dong W, Socher R, *et al.* ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. doi:10.1109/cvprw.2009.5206848

53 Harron K, Dibben C, Boyd J, *et al.* Challenges in administrative data linkage for research. Big Data & Society. 2017;**4**:205395171774567. doi:10.1177/2053951717745678

54 Data protection by design and default. Published Online First: 9 February 2021.https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/ (accessed 13 Jul 2021).

55 Lee AY, Campbell JP, Hwang TS, *et al.* Recommendations for Standardization of Images in Ophthalmology. *Ophthalmology* 2021;**128**:969–70.

56 NHS and social care data: off-shoring and the use of public cloud services. https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/nhs-and-social-care-data-off-shoring-and-the-use-of-public-cloud-services (accessed 13 Jul 2021).

57 Sinha S, Peach G, Poloniecki JD, *et al.* Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes—systematic review and recommendations for reporting. European Journal of Public Health. 2013;**23**:86–92. doi:10.1093/eurpub/cks046

58 Burns EM, Rigby E, Mamidanna R, *et al.* Systematic review of discharge coding accuracy. *J Public Health* 2012;**34**:138–48.

59 Herrett E, Shah AD, Boggon R, *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;**346**:f2350.

60 Wood A, Denholm R, Hollings S, *et al.* Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021;**373**:n826.

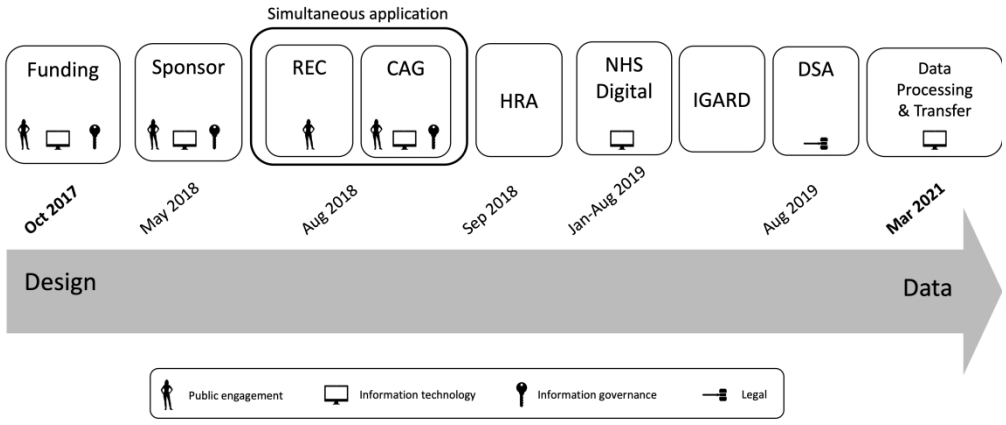


Figure 1: Schematic of the key milestones, prerequisites and approvals with their corresponding achievement dates for the AlzEye dataset. REC: Research Ethics Committee, CAG: Confidential Advisory Group, HRA: Health Research Authority, NHS: National Health Service, IGARD: Independent Group Advising on the Release of Data, DSA: Data Sharing Agreement.

1327x627mm (72 x 72 DPI)

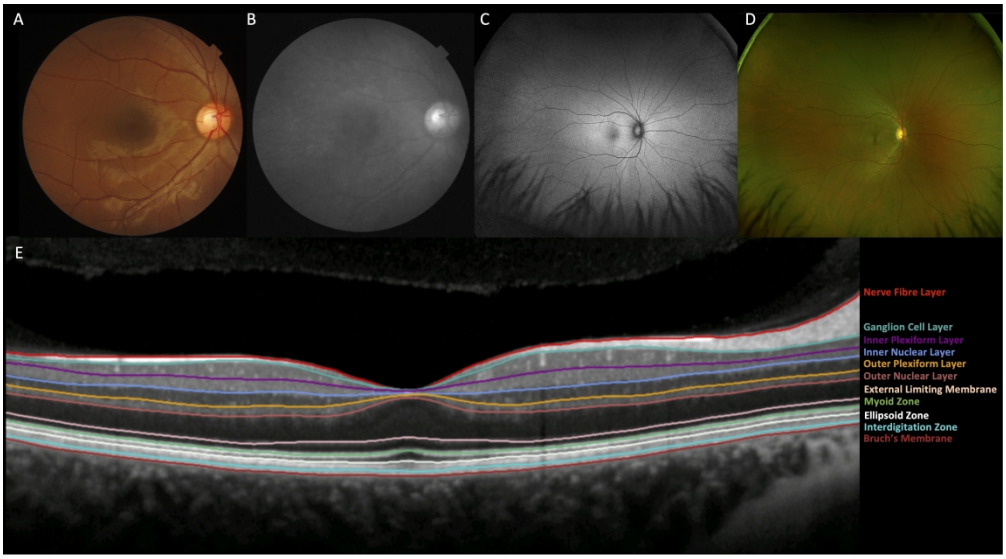


Figure 2: Composite figure showing the major retinal imaging modalities within AlzEye. A: Colour fundus photograph, B: Red-free photograph, C: Fundus autofluorescence (widefield), D: Pseudocolour photography (widefield) and E: Optical coherence tomography of the central macula illustrating segmentation of the individual sublayers. Consensus nomenclature for the retinal sublayers is indicated. NFL= nerve fibre layer, GCL = ganglion cell layer, IPL: inner plexiform layer, INL: inner nuclear layer, OPL: outer plexiform layer, ONL = outer nuclear layer, MZ = myoid zone, RPE-IZ = RPE-interdigitation zone, BM = Bruch's membrane.

1456x803mm (72 x 72 DPI)

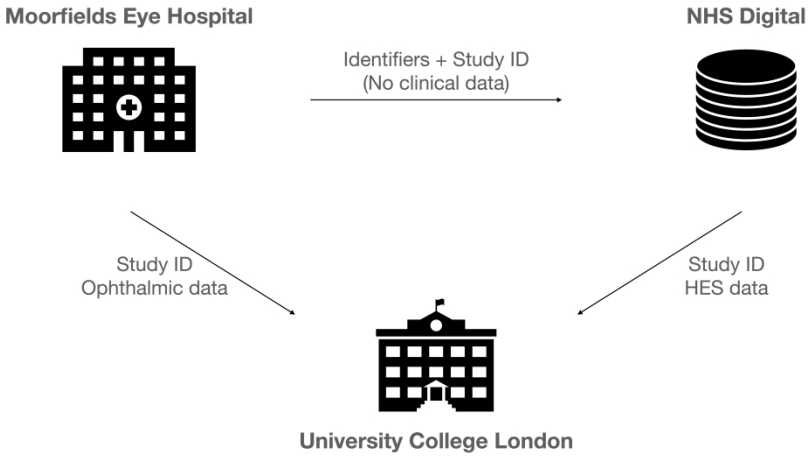


Figure 3: Linkage approach of AlzEye. Moorfields Eye Hospital NHS Foundation Trust securely transfers a spreadsheet of identifiers with a study ID to NHS Digital and separately transfers the study ID with ophthalmic data, including diagnoses and retinal images, to University College London. NHS Digital links the identifiers with the Hospital Episode Statistics (HES) database and returns the admissions data with the study ID (and no identifiable data) to UCL. UCL links the ophthalmic data from Moorfields Eye Hospital with HES data from NHS Digital using the study ID.

2822x1587mm (72 x 72 DPI)

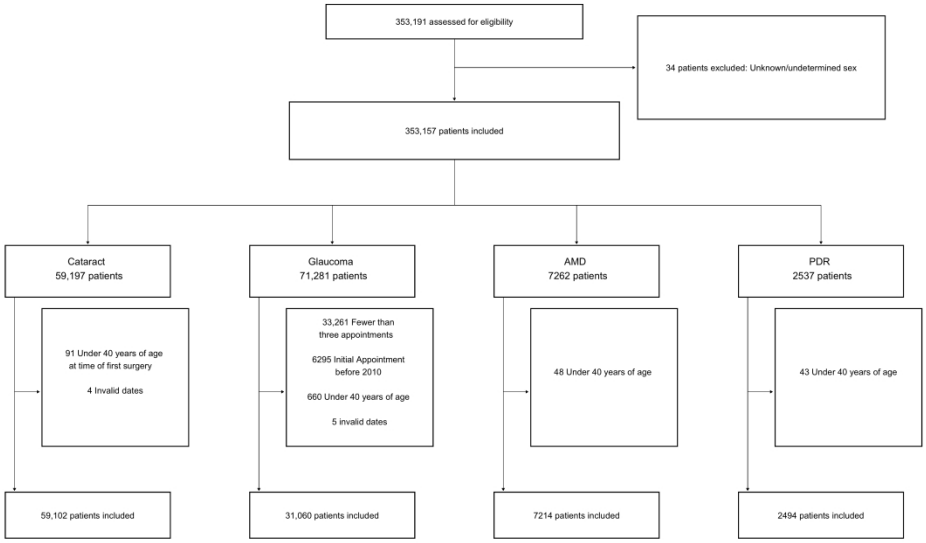


Figure 4: CONSORT style flow chart illustrating the distribution of cataract, glaucoma, neovascular age-related macular degeneration (AMD) and proliferative diabetic retinopathy (PDR) within the AlzEye dataset.

1693x1058mm (72 x 72 DPI)

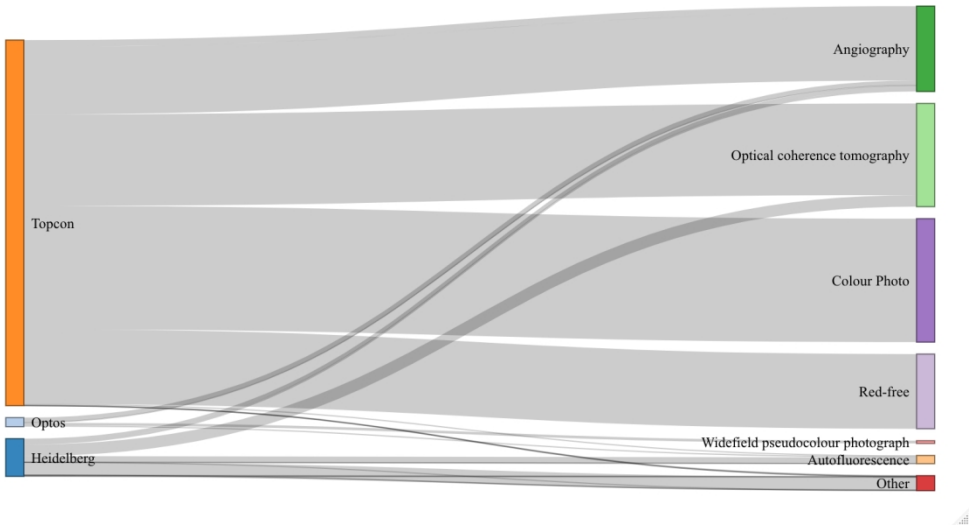


Figure 5: Parallel sets diagram illustrating the imaging modality across vendors within AlzEye. The majority of images were acquired on the Topcon system and the most frequent modalities were colour photography and optical coherence tomography. Designed using the networkD3 package.

584x324mm (72 x 72 DPI)

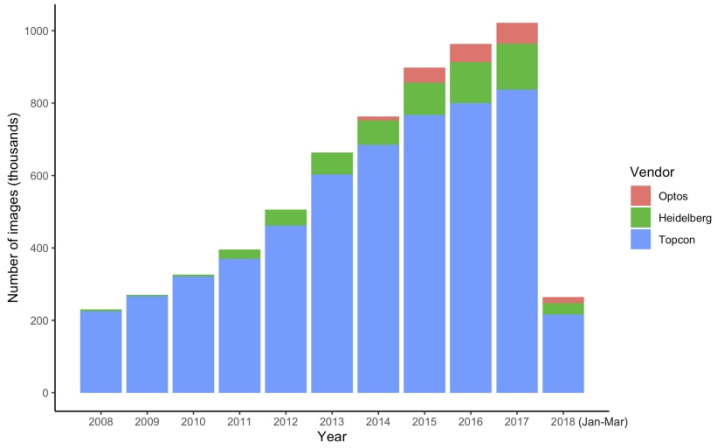


Figure 6: Stacked bar chart of the annual number of images acquired during the study period for the three leading device vendors at Moorfields Eye Hospital. Data for 2018 represents 3 months only prior to the study end-date.

1411x793mm (72 x 72 DPI)

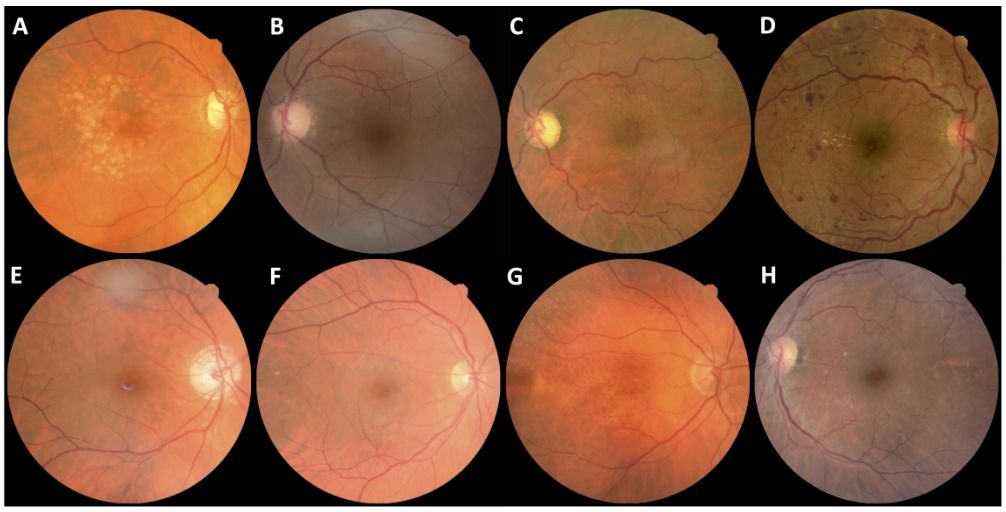


Figure 7: Example colour retinal photographs of ophthalmic and systemic diseases within AlzEye. A: age-related macular degeneration, B: cataract, C: glaucoma, D: proliferative diabetic retinopathy E: prevalent Alzheimer's disease, F: incident ischaemic stroke, G: incident myocardial infarction, H: prevalent vascular dementia.

1800x899mm (72 x 72 DPI)