Strategies for health outcome phenotyping

Cathie Sudlow
UK Biobank Chief Scientist

Ascertaining health outcomes through linking across the UK to NHS datasets covering a wide range of diseases

Data source	Current cohort coverage	Types of data	Coding system
Death registers	100%	Date and cause of death	ICD-10
Cancer registers	100%	Date, stage and grade of cancer	ICD-9 & ICD-10
Hospital admissions	100%	Dates, diagnoses, procedures	ICD-9 & ICD-10 OPCS-4
Primary care	45%	Dates, diagnoses, procedures, symptoms, signs, specialist referrals, prescriptions, lab tests	Read v2 Read v3 BNF / DM&D + others

Maximising the value of the linked healthcare data

Cancers

Cancer register provides high quality disease status information

Non-cancer conditions

Rapid creation of disease status flags for around 1000 conditions based on 3-digit level ICD-10 codes using data from:

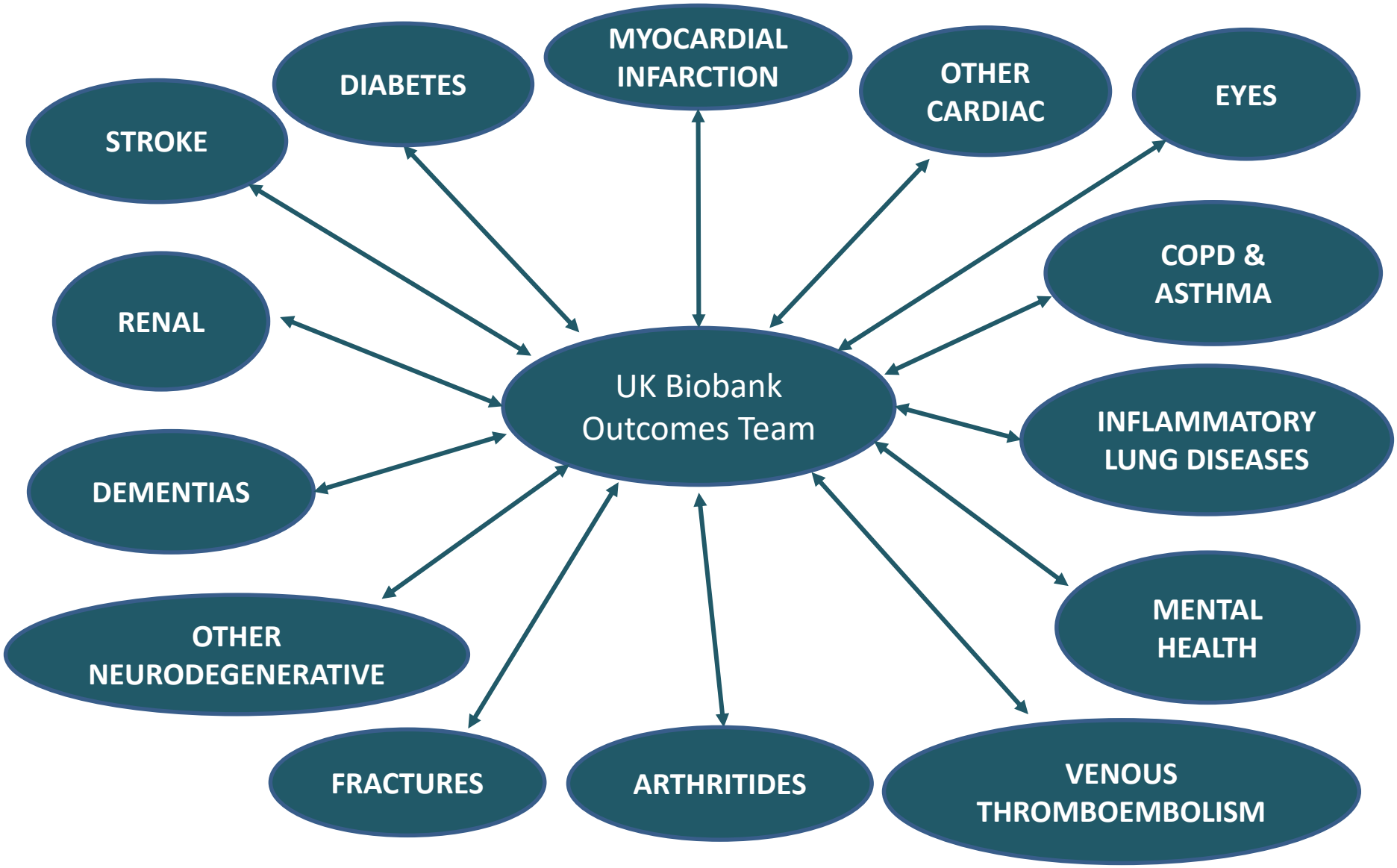
- Self-report at baseline
- Hospital admissions
- Primary care
- Death register

Not perfect but will give coverage of wide range of diseases

Limitations of the linked healthcare data

- Messy ‘real world data’ - not collected for research
- Not 100% accurate - administrative and clinical error
- ‘Comprehensive’ mapping tools to combine data across different coding systems: not perfect, not validated, miss cases
- Some conditions not well captured e.g. mental health, cognition
- Lack information on sub-phenotypes for many conditions
- Hence, UK Biobank working with experts to:
  - Create more accurate disease status indicators
  - Estimate accuracy of these
  - Consider which additional linkages will add most value
  - Scalable approaches for disease sub-classification

Disease status indicators from linked data

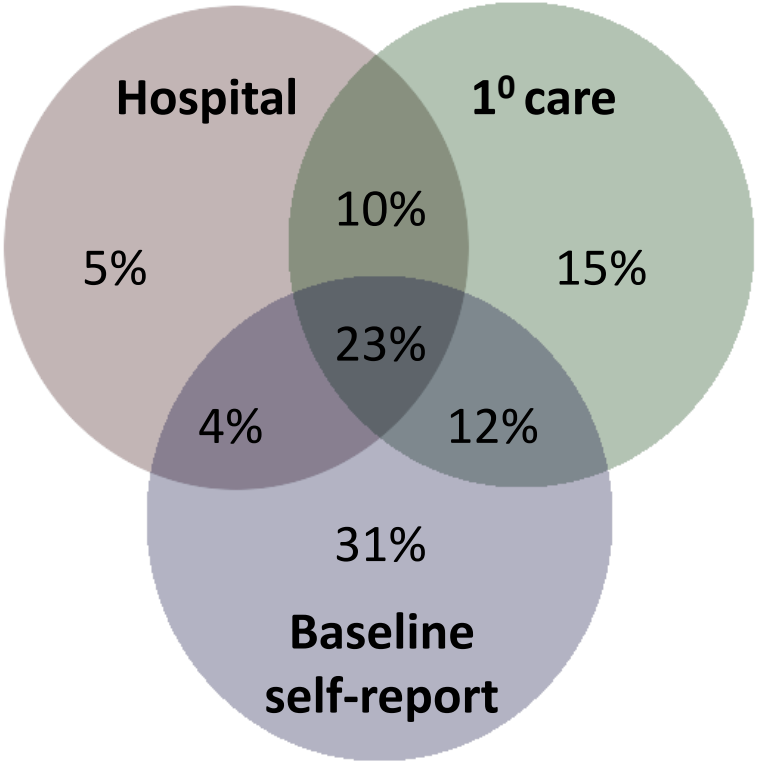


Example of COPD: sources of cases

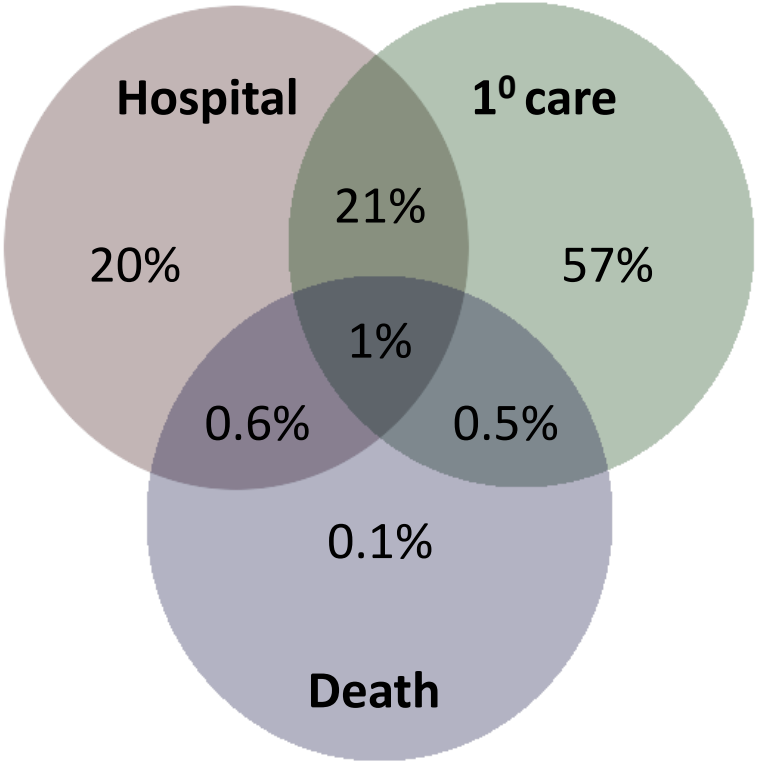
in participants with self-report, hospital, primary care and death data

15 ICD codes; >40 Read v2 codes; >150 Read v3 codes

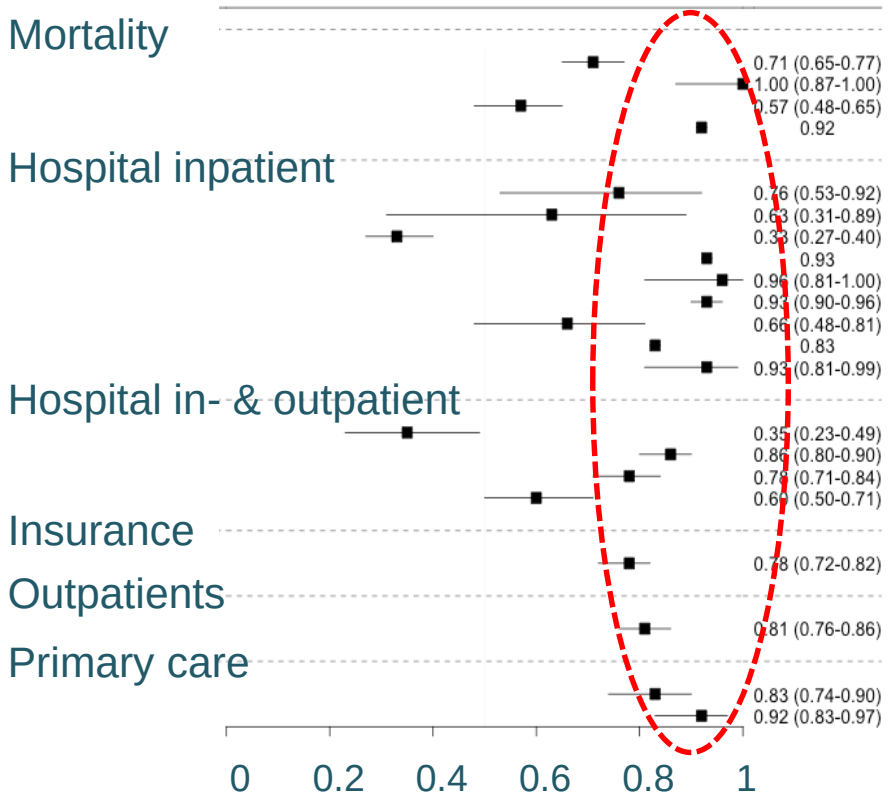
Prevalent cases:
detected before recruitment

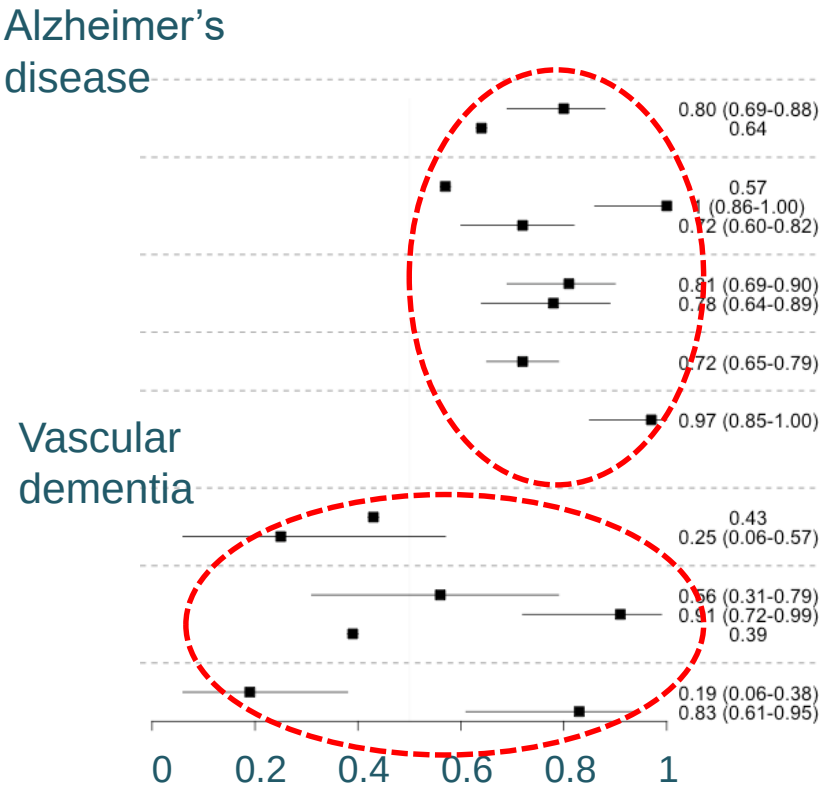


Incident cases:
detected after recruitment



Example of dementia: published studies reporting on accuracy (positive predictive value) of routine health data





Wide variation but in most PPV >80%

PPV for AD higher than for vasc dementia

Assessing accuracy versus expert adjudication of full free text medical records: first results from regional subset of 17,000 participants

120 cases of dementia

80 cases of Parkinson's disease

225 cases of stroke

Assessing accuracy versus expert adjudication of full free text medical records: first results from regional subset of 17,000 participants

	DEMENTIA N=120	PD N=80	STROKE N=225
All codes:			
	83% (75% to 89%)	91% (83% to 96%)	79% (73% to 84%)
Primary care codes:			
	87% (79% to 93%)	95% (87% to 99%)	80% (72% to 86%)
Hospital codes:			
	87% (76% to 95%)	84% (68% to 94%)	89% (82% to 94%)
Death certificate codes:			
	80% (44% to 98%)	86% (42% to 100%)	57% (18% to 90%)

Outcome phenotyping: future plans

- Additional data linkages
  - enhance accuracy
  - ascertain health outcomes not captured currently
  - enable sub-phenotyping
- Additional web questionnaires
  - for health outcomes not captured by linked healthcare data
- Further regional validation studies to assess accuracy of linked data
  - for additional conditions and in additional regions
- ‘Deep dives’ into systems of large hospitals for disease sub-phenotyping, e.g.:
  - pathology reports, digitised images and tumour tissue
  - free text
  - radiology images
- UK Biobank scalable phenotyping data challenge

Funding bodies: