Follow up and adjudication of health outcomes

Prof Cathie Sudlow

Much more than a biobank….

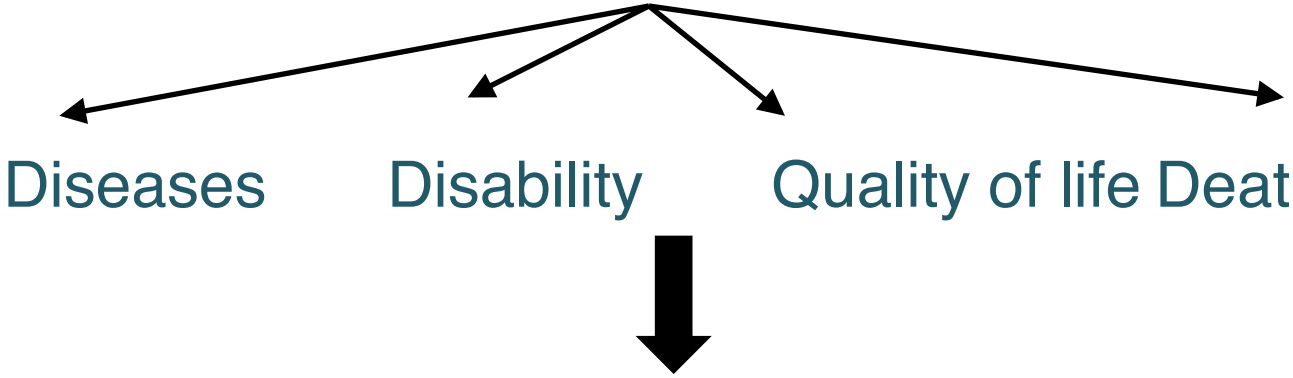
Size +++

Phenotyping +++

Genotyping +++

and

PROSPECTIVE….i.e. follow-up (for decades)



Studies of genetic, lifestyle and environmental determinants of health and disease

Prevalent conditions* at recruitment

Condition	Cases (n)
Diabetes	26,000
MI	12,000
COPD	12,000
Stroke	7,000
Breast cancer	11,000
Colorectal cancer	3,000
Prostate cancer	3,000
Rheumatoid arthritis	6,000
Retinal detachment	2,000

** by self report, confirmed by trained interviewer, rounded to nearest 1000*

Incident outcomes* during follow-up

Condition	2012	2017	2022
Diabetes	10,000	25,000	40,000
MI/CHD death	7,000	17,000	28,000
Stroke	2,000	5,000	9,000
COPD	3,000	8,000	14,000
Breast cancer	2,500	6,000	10,000
Colorectal cancer	1,500	3,500	7,000
Prostate cancer	1,500	3,500	7,000
Hip fracture	1,000	2,500	6,000
Alzheimer's	1,000	3,000	9,000

**Estimates based on UK age- and sex-specific rates, adjusted for potential healthy cohort effects and losses to follow-up , rounded to nearest 500*

Neurodegenerative conditions

Condition	Self-report at baseline (n)	Incident outcomes* (n) by the end of: 2012	2017	2022	2027
Stroke	7,000	2,000	5,000	9,000	20,000
Alzheimer's	100	800	3,000	9,000	30,000
Parkinson's	900	1,000	3,000	6,000	14,000

**Estimates based on UK age- and sex-specific rates, adjusted for potential healthy cohort effects and losses to follow-up, rounded to nearest 1000 (or nearest 100 if < 1000)*

Follow-up of half a million people….

COMPREHENSIVE

SCALABLE

COST EFFECTIVE

All participants:

registered with a GP in the NHS

consented to linkage to health-related records

NHS provides majority of healthcare in UK

National datasets about healthcare & health outcomes exist

….so link to these datasets…..

503,000 participants

22 recruitment centres

89% England

7% Scotland

4% Wales

Strategy for follow-up: data linkage

Key electronic, coded, national sources for all participants:

- Death registrations
- Cancer registrations
- Hospital episode data
- Primary care data*

Data linkage: challenges

- Regulation, bureaucracy, and permissions
- Data transfer
- Matching
- Coding queries
- Mapping between coding systems
- Understanding data structure
- Mapping between UK countries
- Presenting data to researchers

Progress with key cohort-wide linkages

Data type	Country	Data provider	Data with UKB	Available for research
Certified deaths by cause (ICD codes)	England	HSCIC	√	√
	Wales			
	Scotland	Central Register & ISD, NSS	√	√
Registered cancers (ICD codes)	England	HSCIC	√	√
	Wales			
	Scotland	Central Register & ISD, NSS	√	√
Hospital episodes (ICD & procedure codes)	England	HSCIC (HES)	√	√
	Wales	SAIL, Swansea	√	√
	Scotland	ISD (SMR), NSS	√	√
Primary care (Read codes)	England	?	x	x
	Wales	SAIL	√ (50%)	x
	Scotland	HIC, Dundee	√ (50%)	x

Other datasets for linkage:

- Tracing systems
- Enhanced cancer data (NCIN)
- Cancer screening
- Other health datasets: mental health data, diagnostic imaging requests, non cancer screening, dental records, disease registries and audits, dispensing, radiology and laboratory reports..
- Health-related data from other government departments: e.g., DWP (disability and other benefits), HMRC (income) etc…

Adjudication of health outcomes

What do the coded data actually tell us?

How accurate?

How detailed?

How complete?

Do we need to go beyond the coded data?

Outcomes working group

Expert advice on methods for ascertainment, confirmation, and sub-classification of disease outcomes:

- Cancer
- Diabetes
- Cardiac outcomes
- Stroke
- Mental health outcomes
- Ocular outcomes
- Neurodegenerative outcomes
- Respiratory outcomes
- Musculoskeletal outcomes

Value of accurate and detailed phenotyping of outcomes

- Enhanced power
  - through avoiding misclassification
- Increased specificity of disease classification
  - determinants of apparently similar, but aetiologically different, conditions may differ

General principles for outcomes adjudication strategy

Staged approach: ascertain, confirm, sub-classify

Avoid false positive cases (but tolerate some false negatives)

Geographical generalisability

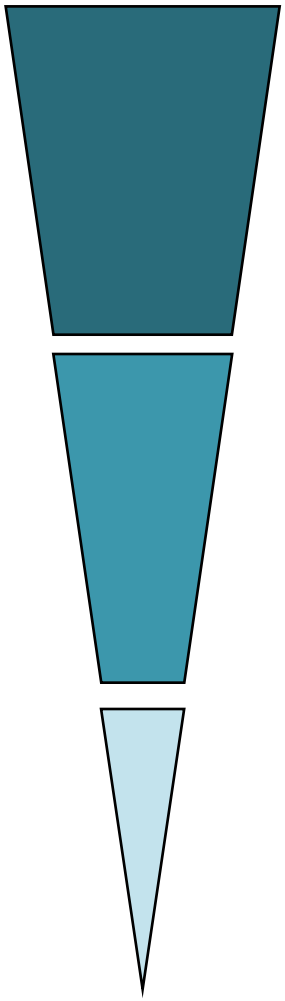
Scalability

Centralised IT development

Cost-effectiveness

Future proof

Staged approaches to outcomes adjudication



APPROACH	CHARACTERISTICS	POSSIBLE DATA SOURCES
ASCERTAINMENT of suspected cases	Cost-effective Feasible Geographically generalisable Scalable	Death registers Cancer registers Hospital episodes (GP records)
CONFIRMATION of "caseness"	As above but somewhat higher cost / lower feasibility	Disease registers Cross-referencing e-records
SUB-CLASSIFICATION of cases	More involved and costly	Tumour collection / assays Specialised databases (eg imaging) Review of clinical records

From linked data to research-ready outcomes: phased implementation

PLANNING

- Scoping, literature reviews, piloting
- Advanced for cancer, MI, stroke, diabetes
- Underway for other areas

PHASE 1

- State-of-the-art coded data – deaths / cancers / hospital episodes
- Algorithms to maximise positive predictive value
- Underway

PHASE 2

- Additional coded data – primary care / enhanced cancer dataset / disease registers
- More complex algorithms

PHASE 3

- 'Expert' adjudication – confirmation and sub-classification of disease
- Incorporate non-coded data from local sources (medical records, imaging etc)
- Interactive web adjudication platform

Funding bodies:

Supported by:

wellcome trust

MRC Medical Research Council

DH Department of Health

British Heart Foundation

DIABETES UK
CARE. CONNECT. CAMPAIGN.

The Scottish Government

Llywodraeth Cymru
Welsh Government

Enhanced power through avoiding misclassification

Minimum detectable odds ratios for association between risk factor and disease in a nested case-control study:

- 80% power
- 10% of controls have the risk factor of interest
- 5000 cases of disease; 4 controls per case

Critical p-value	Misclassification: None*	Some **	More ***
0.01	1.21	1.24	1.28
10^{-7}	1.35	1.43	1.49

** All cases and no controls have disease*
*** 10% cases do not and 2.5% controls do have disease*
**** 20% cases do not and 5% controls do have disease*

Different risk factor profiles for different disease ischaemic stroke subtypes