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

<table>
<thead>
<tr>
<th>Data source</th>
<th>Current cohort coverage</th>
<th>Types of data</th>
<th>Coding system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death registers</td>
<td>100%</td>
<td>Date and cause of death</td>
<td>ICD-10</td>
</tr>
<tr>
<td>Cancer registers</td>
<td>100%</td>
<td>Date, stage and grade of cancer</td>
<td>ICD-9 &amp; ICD-10</td>
</tr>
<tr>
<td>Hospital admissions</td>
<td>100%</td>
<td>Dates, diagnoses, procedures</td>
<td>ICD-9 &amp; ICD-10 OPCS-4</td>
</tr>
<tr>
<td>Primary care</td>
<td>45%</td>
<td>Dates, diagnoses, procedures, symptoms, signs, specialist referrals, prescriptions, lab tests</td>
<td>Read v2, Read v3, BNF / DM&amp;D + others</td>
</tr>
</tbody>
</table>
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:
  o Create more accurate disease status indicators
  o Estimate accuracy of these
  o Consider which additional linkages will add most value
  o Scalable approaches for disease sub-classification
Disease status indicators from linked data

UK Biobank Outcomes Team

- DIABETES
- MYOCARDIAL INFARCTION
- OTHER CARDIAC
- EYES
- COPD & ASTHMA
- INFLAMMATORY LUNG DISEASES
- MENTAL HEALTH
- VENOUS THROMBOEMBOLISM
- FRACTURES
- ARTHRITIDES
- OTHER NEURODEGENERATIVE
- RENAL
- DEMENTIAS
- STROKE
- NEURODEGENERATIVE
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

Hospital: 5%
1st care: 10%
Baseline self-report: 31%

Incident cases: detected after recruitment

Hospital: 20%
1st care: 21%
Death: 0.1%

Prevalent cases detected before recruitment:
- Hospital: 5%
- 1st care: 10%
- Baseline self-report: 31%

Incident cases detected after recruitment:
- Hospital: 20%
- 1st care: 21%
- Death: 0.1%
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

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