Sequencing in cohort studies & large sample collections: Key lessons and reactions

Rory Collins
UK Biobank
Principal Investigator,
Oxford University,
United Kingdom
Future strategies: Reflections on previous experience with GWAS in complex diseases

- Many smaller versus few larger experiments (e.g. need for meta-analyses of many small studies)
- “Deep” versus “shallow” participant phenotyping (e.g. meta-analyses of studies with different definitions)
- Studying few versus many different diseases (e.g. prospective cohorts with extensive record linkage)
- “Deep” versus “shallow” disease phenotyping (e.g. value of ability to characterise disease subtypes)
- National versus international strategic initiatives

What to do in next few years versus what would we want to have done in 10-15 years?
# Approach to disease adjudication: different types of data required at different stages

<table>
<thead>
<tr>
<th>Approach</th>
<th>Characteristics</th>
<th>Possible examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascertainment of suspected cases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmation of “caseness”</td>
<td>As above but somewhat higher cost/lower feasibility</td>
<td>Existing morbidity registers (eg MINAP)</td>
</tr>
<tr>
<td>Classification of confirmed cases</td>
<td>More involved and costly</td>
<td>Targeted blood sampling with costly assays</td>
</tr>
</tbody>
</table>

- **Cost-effective**
- **Feasible**
- **Geographically generalisable**
- **Scalable**

**Possible examples**

- Hospital discharge records
- Primary care records
- Web self-report questionnaire
- Existing morbidity registers (eg MINAP)
- Cross-referencing e-records
- Targeted blood sampling with cheap assays
- Tumour collection/assays
- Specialised databases (eg imaging)
- Review of clinical records
UK Biobank: an international resource

• PROSPECTIVE: It can assess the full effects of a particular exposure (such as smoking) on all types of health outcome (such as cancer, vascular disease, lung disease, dementia)

• DETAILED: The wide range of questions, measures and samples at baseline allows good assessment of exposures, and outcome adjudication allows good disease classification

• LARGE: Inclusion of large number of people allows reliable assessment of the causes of a wide range of diseases, and of the combined impact of many different exposures

• ACCESS: General consent for follow-up through all health records for all types of health research, and for re-contact, by academic and commercial researchers worldwide
UK Biobank: Expected numbers of participants developing diseases during long-term follow-up

<table>
<thead>
<tr>
<th>Condition</th>
<th>2012</th>
<th>2017</th>
<th>2022</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>10,000</td>
<td>25,000</td>
<td>40,000</td>
</tr>
<tr>
<td>MI/CHD death</td>
<td>7,000</td>
<td>17,000</td>
<td>28,000</td>
</tr>
<tr>
<td>Stroke</td>
<td>2,000</td>
<td>5,000</td>
<td>9,000</td>
</tr>
<tr>
<td>COPD</td>
<td>3,000</td>
<td>8,000</td>
<td>14,000</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>2,500</td>
<td>6,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>1,500</td>
<td>3,500</td>
<td>7,000</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>1,500</td>
<td>3,500</td>
<td>7,000</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>800</td>
<td>2,000</td>
<td>4,000</td>
</tr>
<tr>
<td>Hip fracture</td>
<td>800</td>
<td>2,500</td>
<td>6,000</td>
</tr>
<tr>
<td>Rheum. arthritis</td>
<td>800</td>
<td>2,000</td>
<td>3,000</td>
</tr>
<tr>
<td>Alzheimer’s</td>
<td>800</td>
<td>3,000</td>
<td>9,000</td>
</tr>
</tbody>
</table>
Value of detailed phenotyping not only of participants at baseline but also of disease cases during follow-up

- Enhancement of power to detect associations between risk factors and disease outcomes (false positive diagnoses have main adverse impact)

- Increased specificity of disease classification allows the detection of specific associations (e.g. risk factor only linked to disease sub-type)

- “Future-proofing” of the outcome data so that more detailed phenotyping is possible in future (e.g. retain data/samples to allow refined sub-typing)
Advantages of PROSPECTIVE cohorts for studying the causes of different diseases

- Risk factors can be measured before disease develops (helping to avoid “reverse causality”)
- Associations can be assessed with a range of diseases (provided sufficient numbers occur)
- Appropriate controls can be selected from within the same population as the disease cases
- Confounding by other factors is typically less extreme and can be allowed for more fully

But prospective cohorts need to be LARGE
Prospective studies need to be LARGE: CHD versus SBP for 5K vs 50K vs 500K people in the Prospective Studies Collaboration.