Prognosis research: new opportunities in linked electronic health records

Harry Hemingway BA, MB BChir, MSc, FFPH, FRCP
Professor of Clinical Epidemiology
Director, Farr Institute for Health Informatics Research, London

Methods for Evaluating Medical Tests & Biomarkers
15th July 2013 Birmingham
Big (health) data revolution

Chancellor of the Exchequer George Osborne
“world’s best and most complete data-sets in healthcare”
Royal Society
9 November, 2012

Prime Minister David Cameron
to sequence the full genomes of up to 100,000 NHS patients over the next 3-5 years.
“the first country in the world to use DNA codes in the mainstream of the health service.
By unlocking the power of DNA data, the NHS will lead the global race for better tests,
better drugs and above all better care”
10 December 2012

Minister for Science and Universities, David Willetts
“Harnessing 'big data' in the NHS will revolutionise healthcare.
The UK Farr Institute of Health Informatics Research will bring together highly skilled medical and computer scientists, to use electronic health records to improve understanding of a range of diseases.”
3 July 2013
Transformation is in the Air...

"US models are being eclipsed by non-US studies that are much larger, yet considerably less expensive."

"These themes [include] integrating ‘big data’ science into the practice of epidemiology...foster integration with trials."

"A large sample size does not solve every problem. But it allows us to do a lot more, and sometimes it is just what we need."

Mike Lauer, NIH
Disruptive models

According to NIH workshop, June 2013, UK has opportunities to lead with

- UK Biobank
- Electronic health records research
Prognosis Research Strategy (PROGRESS)

Medical Research Council funded international Partnership

**PROGRESS 1: A framework for researching clinical outcomes**

**PROGRESS 2: Prognostic Factor Research**

**PROGRESS 3: Prognostic Model Research**

**PROGRESS 4: Stratified Medicine Research**
Focus of this talk

Cardiovascular disease.....

Global increases in numbers living with

Why rob banks?
What is the prognosis of acute myocardial infarction?

Consider data from
Quality registries (All hospitals, consecutive patients, ongoing)

Linked with
All cause mortality at 30 days

....And compare countries
Cumulative 30-day mortality after AMI

UK
N=391,077
30-day risk of death 10.5%
(95% CI:10.4%, 10.6%)

Sweden
N=119,786
30-day risk of death: 7.6%
(95% CI:7.4%, 7.7%)

National Institute for Cardiovascular Outcomes Research (NICOR), SWEDHEART Chung et al under review 2013
Excess 30 day mortality after admission for heart attack in UK

Robust across
  – strata defined by
    • diagnostic biomarker levels (troponins)
    • other demographic and clinical factors
  – case mix standardised models
  – propensity score matched samples

True mortality excess may be underestimated
  (in UK registry missed cases are more common than in Sweden and have higher mortality).
PROGRESS 1: A framework for researching clinical outcomes

PROGRESS 2: Prognostic Factor Research

PROGRESS 3: Prognostic Model Research

PROGRESS 4: Stratified Medicine Research
Could it be, in part, explained by UK population being more ethnically diverse?
### Predictors of disease onset: a good place to start when considering prognosis

<table>
<thead>
<tr>
<th>Disease</th>
<th>n</th>
<th>Black</th>
<th>Other</th>
<th>South Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Myocardial infarction</strong></td>
<td>12879</td>
<td>0.48 [0.38; 0.61]</td>
<td>0.99 [0.85; 1.16]</td>
<td>1.69 [1.48; 1.92]</td>
</tr>
<tr>
<td><strong>Heart failure</strong></td>
<td>10307</td>
<td>0.86 [0.69; 1.07]</td>
<td>0.99 [0.82; 1.21]</td>
<td>0.87 [0.70; 1.08]</td>
</tr>
<tr>
<td><strong>Peripheral arterial disease</strong></td>
<td>8769</td>
<td>0.63 [0.50; 0.80]</td>
<td>0.72 [0.58; 0.90]</td>
<td>0.66 [0.53; 0.83]</td>
</tr>
</tbody>
</table>

Initial presentation of 12 cardiovascular diseases (3 shown)  
CALIBER 4 linked sources, primary care, hospital, disease registry and death data. Age and sex adjusted. Reference group= white
South Asian ethnicity not prognostic among people with coronary disease

Lower mortality after event in South Asians

Zaman-J et al  Heart 2013
Small study bias in prognostic factor studies

Example: C Reactive protein and outcome of stable coronary disease

N=83 studies

Hemingway et al PloSMed 2010
Study design: literature based systematic review of genetic and other biological markers

Startpoint: neuroblastoma

Endpoint: disease recurrence or mortality

Findings: 130 different markers, **median of one study per marker**

Riley et al *Health Technol Assess* 2003
‘Higher resolution’ enquiry using the scale of electronic health records

Example: haemoglobin threshold

**Startpoint:** 10,000 women with incident stable angina
Primary care electronic health record

**Endpoint:** followed for death and non-fatal MI for 3 years

**Confounders** Adjusted for age, eGFR, risk factors and co-morbidity

**Evidence:** relative risks compared to the lowest risk hb levels

Shah AD et al PloS Med 2011
Medical Research Council funded international Partnership

**PROGRESS 1: A framework for researching clinical outcomes**  

**PROGRESS 2: Prognostic Factor Research**  

**PROGRESS 3: Prognostic Model Research**  

**PROGRESS 4: Stratified Medicine Research**  
Primary care (CPRD)

- Healthy, GP registration
- Stable angina
- Pneumonia hospitalization
- Myocardial infarction hospitalization
- See GP for follow-up
- Death

Timeline:

- New patient check: blood pressure, smoking status, alcohol use etc.
- Diagnosis of stable angina. Blood tests (e.g., cholesterol). Prescription of aspirin, nitrates etc.
- Diagnosis of myocardial infarction
- Blood tests, blood pressure. Prescriptions of beta blocker, statin, ACEi etc.

Hospitalization (HES)

- Admit / discharge dates.
- Primary diagnosis: Viral pneumonia, not elsewhere classified
- Admit / discharge dates.
- Primary diagnosis: Acute myocardial infarction
- Procedure: Percutaneous coronary intervention

Disease Registry (MINAP)

- ECG, cardiac markers.
- Diagnosis: STEMI

Death Census (ONS)

- Date of death. Cause: 1) Rupture of abdominal aortic aneurysm
- 2) Old myocardial infarction

Linked health records data: achievable?

- N people = all whole populations $10^6$
- Ages = all womb to tomb
- N drugs = all pharmome
- Biomarker values = all clinically measured
- N Diseases = all diseasesome $10^4$
Outcomes assessment: importance of linking multiple record sources

Example: acute myocardial infarction

Death registry

Disease registry

Hospital admissions

Primary care

All four data sources

Crude annual incidence of myocardial infarction per 100 000

Herrett, Shah et al CALIBER    BMJ 2013
Multiple record linkages

Are not ‘databases’

Data are not ‘routine’

They are ‘big data’

In the sense that they need tools for scalable research
Some tools are emerging - CALIBER
How to define phenotypes using multiple EHR data sources?
AF algorithm

1. **Patient**
   - **AF diagnosis in GPRD or HES?**
     - **YES**
       - **Warfarin prescription?**
         - **NO**
           - **Heart valve replacement before prescription?**
             - **YES**
               - **Digoxin prescription?**
                 - **YES**
                   - **Not AF case**
                 - **NO**
                   - **DVT/PE event before prescription?**
                     - **YES**
                       **Inferred**
                     - **NO**
                       **Not AF case**
             - **NO**
               - **Heart failure before prescription?**
                 - **YES**
                   **Inferred**
                 - **NO**
                   **Not AF case**
     - **NO**
       - **History or monitoring terms in record?**
         - **YES**
           **Inferred**
         - **NO**
           **Recorded prior to diagnosis code?**
             - **YES**
               **Inferred**
             - **NO**
               **Historical diagnosis**
   - **NO**
     - **Only history or monitoring terms in record?**
       - **YES**
         **Inferred**
       - **NO**
         **Recorded prior to diagnosis code?**
           - **YES**
             **Inferred**
           - **NO**
             **Historical diagnosis**
### Atrial Fibrillation

**Name**  
af

**Chapter**  
Circulatory disease/Atrial fibrillation

**Definition**  
Diagnosis of atrial fibrillation.

**Data Type**  
Categorical

**Data sources**  
GPRD, HES

**Dictionaries**  
Read, ICD10, BNF, Free text

**Authors**  
K. Morley (UCL), Shah A. (UCL), Patel R. (UCL), Liam Smeeth (LSHTM), R. Schilling (St Bartholomews & The Royal London Hospital), R. Hunter (St Bartholomews & The Royal London Hospital)

**Agreed**  
01/02/2013 (Revision 1)

**Category**  
Definition

#### Source variables

<table>
<thead>
<tr>
<th>Description</th>
<th>Source</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial fibrillation diagnosis</td>
<td>Primary care</td>
<td>at_gprd</td>
</tr>
<tr>
<td>Atrial fibrillation diagnosis</td>
<td>Secondary care</td>
<td>af_hes</td>
</tr>
<tr>
<td>Atrial fibrillation procedures</td>
<td>Primary care</td>
<td>af_proc_gprd</td>
</tr>
<tr>
<td>Atrial fibrillation procedures</td>
<td>Secondary care</td>
<td>af_proc_opcs</td>
</tr>
<tr>
<td>AF medication</td>
<td>Primary care</td>
<td>at_drugs_gprd</td>
</tr>
<tr>
<td>warfarin or digoxin prescription</td>
<td>Primary care</td>
<td>at_warfarin_digoxin</td>
</tr>
<tr>
<td>Deep vein thrombosis</td>
<td>Primary care</td>
<td>dvt_gprd</td>
</tr>
<tr>
<td>Deep vein thrombosis</td>
<td>Secondary care</td>
<td>dvt_hes</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td>Primary care</td>
<td>pe_gprd</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td>Secondary care</td>
<td>pe_hes</td>
</tr>
<tr>
<td>ECG Text/Notes text mining</td>
<td>Secondary care</td>
<td>Algorithm</td>
</tr>
</tbody>
</table>
Startpoint: stable coronary disease (n=102,023)
Endpoint: coronary death and non-fatal MI (n=8,856) over five years

Rapsomaniki et al CALIBER SCAD Prognostic model, under review 2013
Prognostic factors in CALIBER stable CAD model for non-fatal MI / fatal CHD*

<table>
<thead>
<tr>
<th>SCAD DIAGNOSIS &amp; SEVERITY</th>
<th>HR</th>
<th>CVD CO-MORBIDITIES</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other CHD vs. stable angina</td>
<td>1.18</td>
<td>Heart failure, yes vs. no</td>
<td>1.18</td>
</tr>
<tr>
<td>Unstable angina vs. stable angina</td>
<td>1.32</td>
<td>Peripheral arterial disease, yes vs. no</td>
<td>1.08</td>
</tr>
<tr>
<td>NSTEMI vs. stable angina</td>
<td>1.94</td>
<td>Atrial fibrillation, yes vs. no</td>
<td>0.95</td>
</tr>
<tr>
<td>STEMI vs. stable angina</td>
<td>2.37</td>
<td>Stroke, yes vs. no</td>
<td>1.13</td>
</tr>
<tr>
<td>PCI in last 6 months, yes vs. no</td>
<td>0.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG in last 6 months, yes vs. no</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous/recurrent MI, yes vs. no</td>
<td>1.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of long-acting nitrates, yes vs. no</td>
<td>1.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CVD RISK FACTORS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex-smoker vs. never</td>
<td>1.09</td>
</tr>
<tr>
<td>Current smoker vs. never</td>
<td>1.21</td>
</tr>
<tr>
<td>Hypertension, yes vs. no</td>
<td>1.06</td>
</tr>
<tr>
<td>Diabetes mellitus, yes vs. no</td>
<td>1.38</td>
</tr>
<tr>
<td>TCHOL, per 1 mmol/L increase</td>
<td>1.06</td>
</tr>
<tr>
<td>HDL, per 0.5 mmol/L increase</td>
<td>0.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NON-CVD COMORBIDITIES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic kidney disease, yes vs. no</td>
<td>1.08</td>
</tr>
<tr>
<td>COPD, yes vs. no</td>
<td>1.18</td>
</tr>
<tr>
<td>Cancer, yes vs. no</td>
<td>0.95</td>
</tr>
<tr>
<td>Chronic liver disease, yes vs. no</td>
<td>1.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PSYCHOSOCIAL CHARACTERISTICS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression at diagnosis, yes vs. no</td>
<td>1.05</td>
</tr>
<tr>
<td>Anxiety at diagnosis, yes vs. no</td>
<td>1.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BIOMARKERS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart rate, per 10 beats/min increase</td>
<td>1.06</td>
</tr>
<tr>
<td>Creatinine, per 30 μmol/L increase</td>
<td>1.08</td>
</tr>
<tr>
<td>WCC per 1.5 10^9/L increase</td>
<td>1.11</td>
</tr>
<tr>
<td>Haemoglobin, per 1.5 g/dL increase</td>
<td>0.82</td>
</tr>
</tbody>
</table>

*Additionally adjusted for age, sex and social deprivation

Rapsomaniki et al CALIBER SCAD Prognostic model, under review 2013
Prognostic models in stable CAD: Which factors add to discrimination?

Characteristics used in risk-assessment (incrementally updated models)

1. **Sociodemographics**
   - age, sex, deprivation

2. **CAD Diagnosis & Severity**
   - CAD subtype, PCI/CABG, hist MI, nitrate use

3. **Primary CVD Risk Factors**
   - smoking, hypertension, diabetes, lipids

4. **CVD Co-morbidities**
   - heart failure, PAD, atrial fibrillation, stroke

5. **Non-CVD Comorbidities**
   - chronic renal disease, COPD, cancer, liver disease

6. **Psychosocial**
   - depression, anxiety

7. **Biomarkers**
   - heart rate, creatinine, white cell count, haemoglobin

Rapsomaniki et al CALIBER SCAD Prognostic model, under review 2013
...modelling impact
Limitations of EHR approaches

Data quality

Imaging results

Rear view mirror

...complementary approaches with bespoke, investigator led clinical cohorts
Potential advantages

Intrinsic clinical relevance

Decision support implementable in EHR systems (e.g. Q risk)

If ‘diseaseome’ available > opportunities for flexibility in startpoint and endpoint selection

If ‘all’ patients included > what role for ‘external validation’?
## Comparison of stable CAD prognostic models

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size, N</td>
<td>8557</td>
<td>3031</td>
<td>7311</td>
<td>102 023</td>
</tr>
<tr>
<td>Population based</td>
<td>No (trial)</td>
<td>No (voluntary)</td>
<td>No (trial)</td>
<td>Yes</td>
</tr>
<tr>
<td>Startpoint:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stable angina</td>
<td>0%</td>
<td>95%</td>
<td>50%</td>
<td>66%</td>
</tr>
<tr>
<td>Stable after ACS</td>
<td>100%</td>
<td>5%</td>
<td>50%</td>
<td>34%</td>
</tr>
<tr>
<td>Endpoint:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fatal CHD/nfMI</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>Yes</td>
</tr>
<tr>
<td>Endpoint, N</td>
<td>1190</td>
<td>328</td>
<td>1063</td>
<td>8 856 nfMI/fatal CHD 20 817 deaths</td>
</tr>
<tr>
<td>Clinical history</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biomarkers (continuous)</td>
<td>n=2 TCHOL, HDL</td>
<td>n=0</td>
<td>n=4 SBP, CREAT, FPG, WCC</td>
<td>n=6 HR, TCHOL, HDL, CREAT, WCC, Hb</td>
</tr>
<tr>
<td>Externally validated</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Yes</td>
</tr>
<tr>
<td>Used clinical records</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Prognosis Research Strategy (PROGRESS)

Medical Research Council funded international Partnership

**PROGRESS 1: A framework for researching clinical outcomes**

**PROGRESS 2: Prognostic Factor Research**

**PROGRESS 3: Prognostic Model Research**

**PROGRESS 4: Stratified Medicine Research**
Embedding genomics in the EHR:

Example the eMERGE Consortium

Figure 2. Phenome-wide association study plots of the most significant single-nucleotide polymorphisms associated with QRS duration. A, SCN5A (rs1805126). B, SCN10A (rs6795970). Blue lines indicate $P < 0.05$.}

Ritchie et al Circulation 2013
Conclusion

Prognostic studies are important design in evaluating medical tests and biomarkers

Linkage of multiple sources of electronic health record growing internationally

Use across four PROGRESS themes in prognosis research becoming more established
‘An arsenal that the genius of English healers cannot fail to turn to account’

William Farr 1874

supplement to 35th annual report of the Registrar General, page lxxx