i2b2, Cafe Variome and tranSMART: The Forefront of Health Data Management and Discovery in Leicester

Tim Beck and Adam Webb
Department of Genetics and Genome Biology
Overview

• Biomedical Informatics in Leicester

• i2b2 and Cafe Variome
  – Example: The Genetics and Vascular Health Check study (GENVASC)

• tranSMART
  – Example: The COPDMAP project
Leicester's Biomedical Informatics Network for Education, Research and Industry (BINERI)

- Interdisciplinary grouping to bring together expertise in biomedical informatics, healthcare data management, and information technology
- Unifies activities across Leicester concerned with:
  - Data science
  - Bioinformatics training
  - Data discovery and sharing
  - Biobanking
  - Big data analysis
  - Governance
  - Ethics
  - Patient engagement
healthcare
warehouses
e.g., i2b2,
tranSMART,
LabKey
tissue tracking
e.g.,
OpenSpecimen
patient relationship management
e.g.,
CiviCRM
CDISC standards
healthcare
data warehouses
e.g., i2b2,
tranSMART, LabKey
eCRF platforms
e.g., REDCap
data discovery
e.g., Cafe Variome
BINERI
Leicester
Leicester Cardiovascular BRU
Leicester Respiratory BRU
Leicester-Loughborough Diet, Lifestyle and Physical Activity BRU
NIHR Collaborative Leadership in Applied Health Research & Care
Cancer Research UK Leicester Centre
University Hospitals of Leicester NHS Trust
Leicester Clinical Trials Unit
Leicester Precision Medicine Institute
Leicester College of Science and Engineering
Leicester College of Medicine, Biological Sciences and Psychology
University of Leicester IT Services
Leicester Respiratory BRU
Leicester BRC
stakeholders
domain expertise
The **GENVASC** study

- Add genetic screening to NHS Health Check
- Recruitment & blood sample within GP surgery

Questions GENVASC will be answering:

By what mechanisms do CAD-associated loci affect coronary risk?

Can a genetic risk score improve CAD risk prediction and primary prevention?

Can CAD be reliably diagnosed by a blood test?

Can we improve our understanding of rarer cardiovascular diseases?
The GENVASC study

• 117 recruiting practices
• >24,000 participants recruited by GPs in 4 years
• 169 participants have experienced 205 cardiac episodes since recruitment
• Track primary care records for 15 years
Genetic Queries
Share the ‘existence’ rather than the ‘substance’ of data

This technology (or similar) sits atop/alongside existing local DBs to bring the discoverability and connectivity, without replacing or altering the local solutions

www.cafevariome.org
Cafe Variome Discovery

Query Builder

Cafe Variome Federated

Access Control

User, group and record access control management

<table>
<thead>
<tr>
<th>Source</th>
<th>openAccess</th>
<th>linkedAccess</th>
<th>restrictedAccess</th>
</tr>
</thead>
<tbody>
<tr>
<td>mockdata_1 (Cafe Variome Demo 4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mockdata_2 (Cafe Variome Demo 4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mockdata_3 (Cafe Variome Demo 4)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Collaborating Networks
LCB Cafe Variome

- Allow discovery across integrated genotype (VCF) and phenotype/demographic (i2b2 deposited) data

- Enable cohorting by Leicester researchers (display participant IDs)

- Enable authorised users to identify how many Leicester participants match a specific query (display counts only), e.g.
  - how many caucasian participants aged between 40 and 90 have a history of aortic stenosis and a homozygous variant at SNP rs10455872?
LCB Cafe Variome

GENOTYPE: Pseudonymised VCF file

PHENOTYPE: Pseudonymised export file

GENVASC i2b2

University

Hospital
Cafe Variome Data and Index Models

Data Model: MySQL (relational tables)

Index Model: Elasticsearch (JSON)

One entity
One distinct core attribute
One distinct Attribute:Value
<table>
<thead>
<tr>
<th>record_id</th>
<th>source</th>
<th>DOB</th>
<th>Gender</th>
<th>Visit[year]</th>
<th>WC</th>
<th>WC[unit]</th>
<th>Cognitive disorder/latest diagnosis</th>
<th>MMSE score</th>
<th>Follow up length</th>
<th>Follow up[unit]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFHS:3629385</td>
<td>GenerationS1</td>
<td>1949</td>
<td>Female</td>
<td>2006</td>
<td>77.3 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>12 month</td>
</tr>
<tr>
<td>SFHS:3629386</td>
<td>GenerationS1</td>
<td>1961</td>
<td>Female</td>
<td>2008</td>
<td>81.3 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>12 month</td>
</tr>
<tr>
<td>SFHS:3629387</td>
<td>GenerationS1</td>
<td>1976</td>
<td>Female</td>
<td>2010</td>
<td>81.8 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>13 month</td>
</tr>
<tr>
<td>SFHS:3629388</td>
<td>GenerationS1</td>
<td>1948</td>
<td>Female</td>
<td>2009</td>
<td>79.3 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>12 month</td>
</tr>
<tr>
<td>SFHS:3629389</td>
<td>GenerationS1</td>
<td>1951</td>
<td>Female</td>
<td>2008</td>
<td>75.6 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>18 month</td>
</tr>
<tr>
<td>SFHS:3629390</td>
<td>GenerationS1</td>
<td>1947</td>
<td>Male</td>
<td>2010</td>
<td>90.6 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>24 month</td>
</tr>
<tr>
<td>SFHS:3629391</td>
<td>GenerationS1</td>
<td>1975</td>
<td>Female</td>
<td>2010</td>
<td>74.1 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>13 month</td>
</tr>
<tr>
<td>SFHS:3629392</td>
<td>GenerationS1</td>
<td>1970</td>
<td>Male</td>
<td>2010</td>
<td>96.5 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>24 month</td>
</tr>
<tr>
<td>SFHS:3629393</td>
<td>GenerationS1</td>
<td>1953</td>
<td>Male</td>
<td>2007</td>
<td>91.4 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>16 month</td>
</tr>
<tr>
<td>SFHS:3629394</td>
<td>GenerationS1</td>
<td>1957</td>
<td>Female</td>
<td>2006</td>
<td>80.4 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>12 month</td>
</tr>
<tr>
<td>SFHS:3629395</td>
<td>GenerationS1</td>
<td>1934</td>
<td>Female</td>
<td>2009</td>
<td>73.5 cm</td>
<td>cm</td>
<td>normal</td>
<td></td>
<td>Unknown</td>
<td>13 month</td>
</tr>
</tbody>
</table>

**EAV NEST**

- **Attribute:** A
- **Value:** V
- **Date_From:** DF
- **Date_To:** Dt
- **Meta:** M
• £6m funding from MRC/ABPI over 4 years
• COPD Deep Phenotyping
• Mechanisms, impact and therapeutic targeting of **microbial and viral colonisation** in COPD
• Tissue repair and injury
• Reducing the burden of COPD by targeting skeletal muscle mass and function. Targets and endpoints for drug development
50+ low-dimension datasets

- 525 patients
- Clinical data
- Lab tests
- Many generated at stable & exacerbation

Datasets – derived from blood and sputum

- Genomics (linking) (blood)
- Microbiomics (stable & exacerbation) (sputum)
- Transcriptomics (RNAseq) (blood)
- qPCR (sputum)
Complex visit structure

- Regular visit schedule gets interrupted by exacerbations

- Baseline
- 6 monthly
- 12 monthly
- Exacerbation, plus follow-ups

525 patients
Up to 6 stable visits \textit{plus}
0-9 exacerbation visits \textit{plus}
2w + 6w follow-ups
Design → Data collection → Analysis
Design

Data collection

Analysis
COPDMAP Data Flow

- OpenClinica Clinical phenotypes database
- Sample tracking system
- BioMart clinical phenotype data output
- Lab local data system
- Lab-based research site
- Lab
- LabKey Data integration platform
- tranSMART Data integration platform
- Recruitment site
- CRF data
- Sample data
- COPDMAP central components
- ETL
- ETL & mapping
- Quality controller (Exploristics)
LabKey

• PROs
  – Intuitive approach. Displays all data in organised sets and allows filtering and sorting.
  – Holds additional data which cannot be loaded into tranSMART (medications, full text, etc).
  – Flexible user access controls (datasets and cohorts)
  – Basic sample information
  – R API

• CONs
  – Difficult to build complex queries and custom exports
  – Limited built-in analyses
  – Limited support for multi-dimensional data (OMICs etc)
tranSMART

• Why tranSMART?
  – Widespread use and support
  – Rapid generation of data summaries
  – More complex analyses
  – Multi-OMICs capabilities
  – Aligning COPD MAP data with other studies

  – Workflow based approach.
    • Define cohort -> Summary data -> Analyse / Export

• Problems?
  – Lengthy ETL process
  – Limited fine-grained access control (by ontology nodes, not by subject)
  – Complex visit structure
  – Some data types not supported (e.g. full text, detailed medications, dates)
COPDMAP
tranSMART tree

ETL:
Baseline: Pfizer
Baseline, follow-up, exacerbation: Serge Eifes (ITTM)
Multiple visits in tranSMART

- Each variable is sub-divided
- Every variable / visit combination becomes a variable

- Example:
  - FEV1 collected at 0, 3, 6, 12, 18, 24, 30, 36m
  - 7 exacerbations (101..107)
  - 4 exacerbation + 2 week (101.2 ... 104.2)
  - 4 exacerbation + 6 week (101.6 ... 104.6)

- Some data can’t be included in tranSMART
  - Free-text fields
  - Dates (get converted to strings) – use day offsets
  - Data are still available in LabKey
Acknowledgements

Bioinformatics Research Group
www.le.ac.uk/bioinformatics

Group leader
Prof Anthony Brookes

Cafe Variome project lead
Dr Colin Veal

Senior developers
Dhiwa Thangavelu
Dr Owen Lancaster

Leicester Biomedical Research Centre: Cardiovascular Theme

Informatics team
Richard Bramley
Daniel Lawday
Susan Sterland
Shajid Issa

BioResource manager
Dr Gavin Whyman