

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



UNIVERSITY OF
LEICESTER

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

Leicester BRC



BINERI
Leicester



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 ?

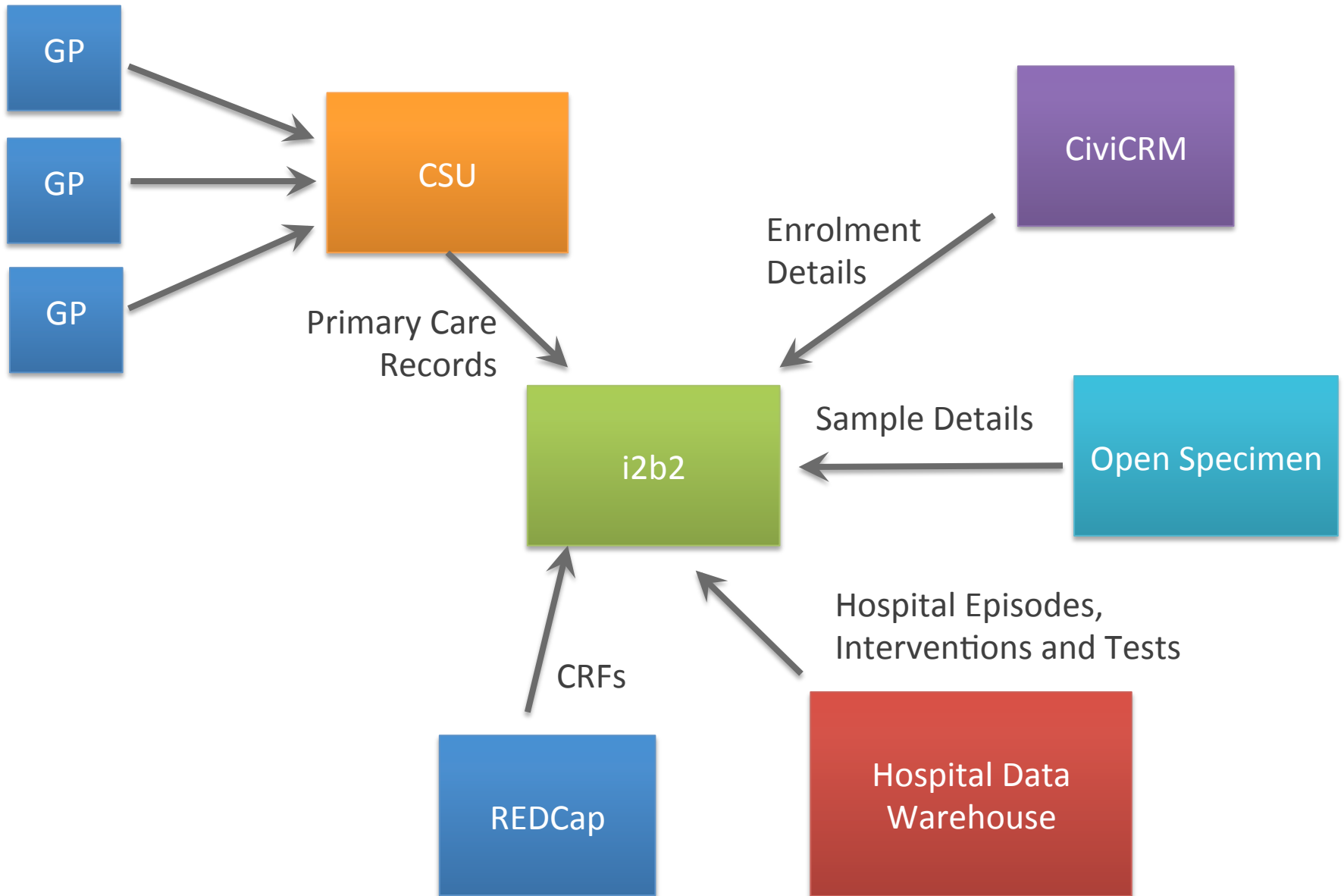
Can CAD be reliably diagnosed by a blood test?

Can we improve our understanding of rarer cardiovascular diseases?

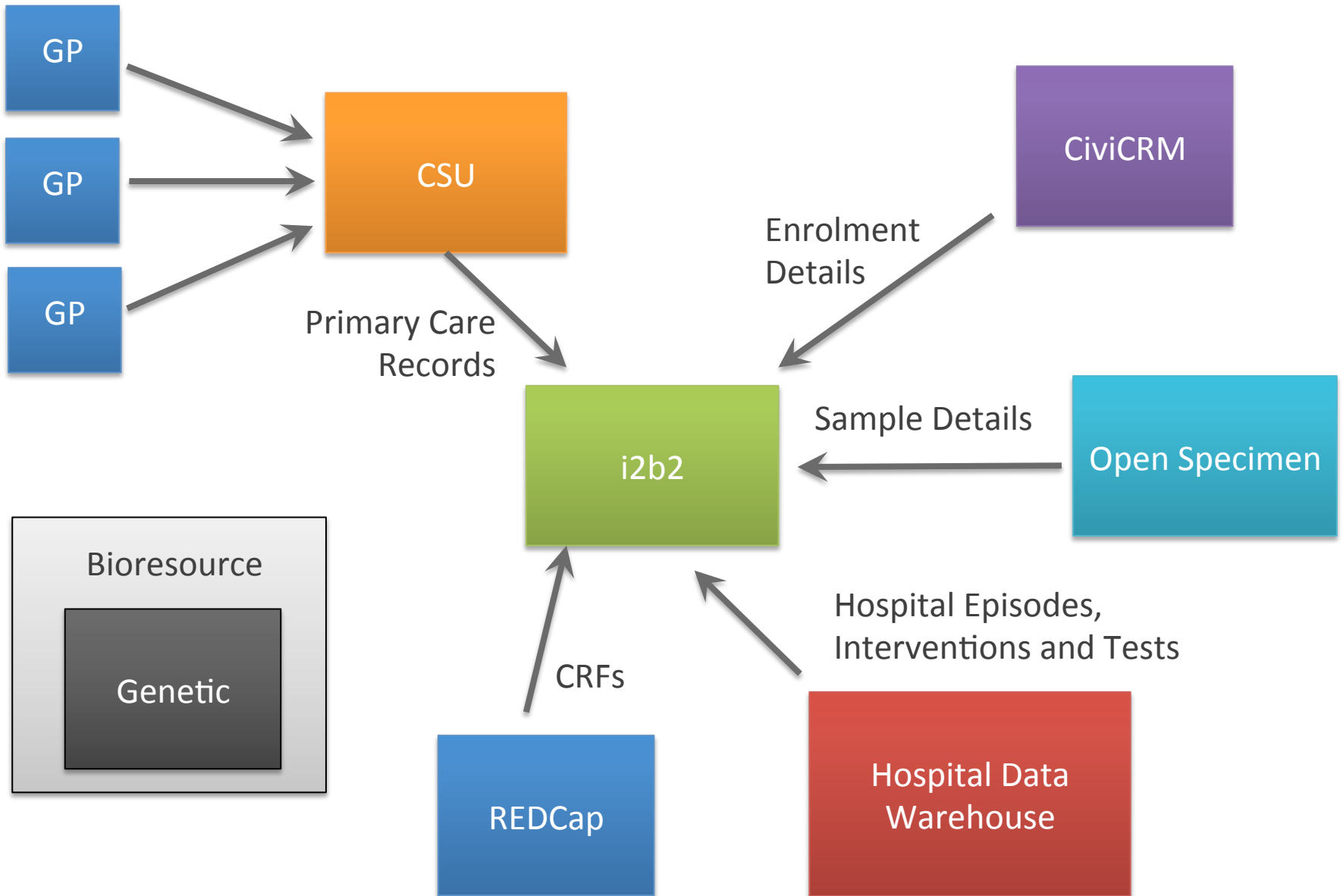
The **GENVASC** study

- 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

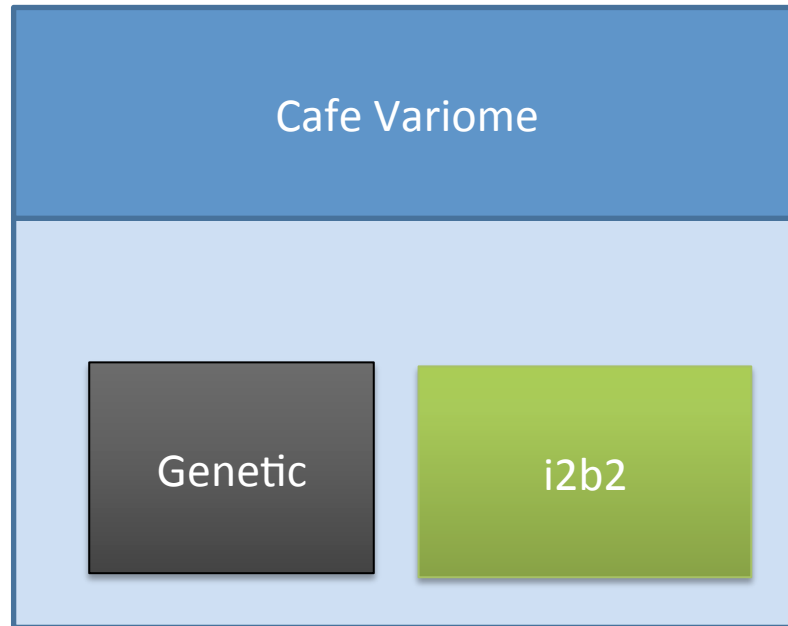
GENVASC Data



GENVASC Data



Genetic Queries





Cafe Variome

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

Query Builder

GENOME COORDINATE AND OR

ACCESSION COORDINATE AND OR

DNA SEQUENCE OF VARIANT AND OR

PROTEIN SEQUENCE OF VARIANT AND OR

GENE SYMBOL AND OR

HGVIS NAME AND OR

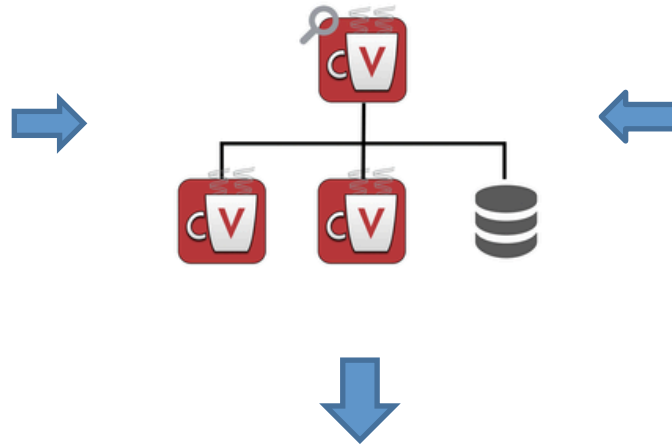
PHENOTYPE

Gender IS | male

AND OR

OTHER SEARCH FIELDS

Cafe Variome Federated



Access Control



((Gender:male))

Source	openAccess		linkedAccess		restrictedAccess	
mockdata_1 (Cafe Variome Demo 4)						
mockdata_2 (Cafe Variome Demo 4)	8		0		0	
mockdata_3 (Cafe Variome Demo 4)						

Collaborating Networks



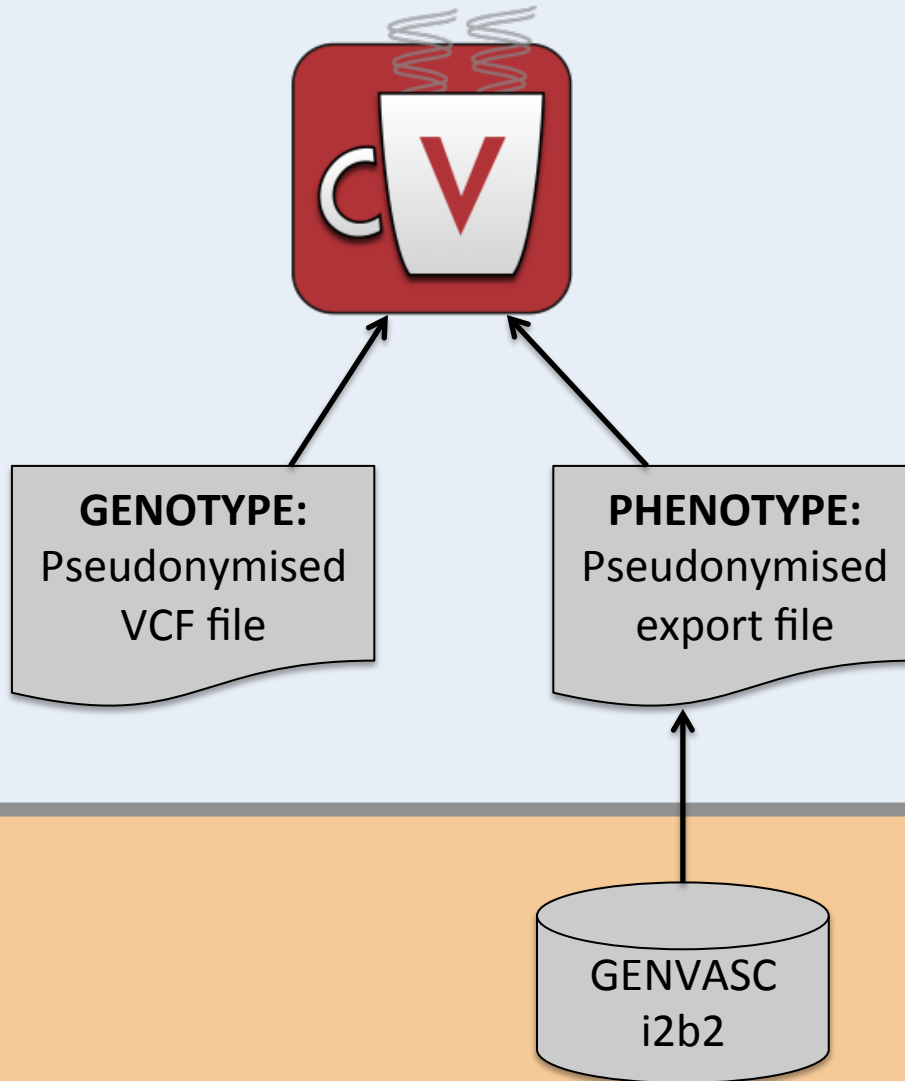
LCB Cafe Variome



- 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

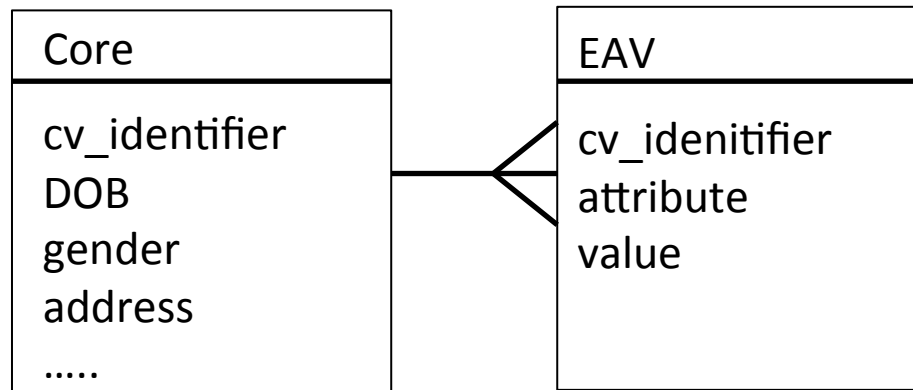
University



Hospital

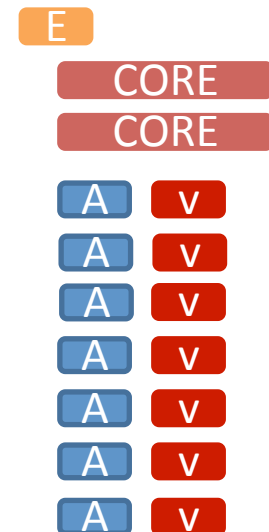
Cafe Variome Data and Index Models

Data Model : MySQL (relational tables)



Index Model : Elasticsearch (JSON)

One entity
One distinct core attribute
One distinct Attribute:Value



	A	B	C	D	M	A	M	A	A	A	M
1	record_id	source	DOB	Gender	Visit[year]	WC	WC[unit]	Cognitive disorder[latest diagnosis]	MMSE score	Follow up length	Follow up[unit]
2	SFHS:3629385	GenerationSi	1949	Female	2006	77.3	cm	normal	Unknown	12	month
3	SFHS:3629386	GenerationSi	1961	Female	2008	81.3	cm	normal	Unknown	12	month
4	SFHS:3629387	GenerationSi	1976	Female	2010	81.8	cm	normal	Unknown	13	month
5	SFHS:3629388	GenerationSi	1948	Female	2009	79.3	cm	normal	Unknown	12	month
6	SFHS:3629389	GenerationSi	1951	Female	2008	75.6	cm	normal	Unknown	18	month
7	SFHS:3629390	GenerationSi	1947	Male	2010	90.6	cm	normal	Unknown	24	month
8	SFHS:3629391	GenerationSi	1975	Female	2010	74.1	cm	normal	Unknown	13	month
9	SFHS:3629392	GenerationSi	1970	Male	2010	96.5	cm	normal	Unknown	24	month
10	SFHS:3629393	GenerationSi	1953	Male	2007	91.4	cm	normal	Unknown	16	month
11	SFHS:3629394	GenerationSi	1957	Female	2006	80.4	cm	normal	Unknown	12	month
12	SFHS:3629395	GenerationSi	1934	Female	2009	73.5	cm	normal	Unknown	13	month

CORE

Core table

E CORE

E

CORE

CORE

EAV NEST

Attribute: A

Value: V

Date_From: DF

Date_To: Dt

Meta: M

Attribute: A

Value: V

Date_From: DF

Date_To: Dt

Meta: M

EAV

E A V DF Dt M

V

visit_dimension		
PK	<u>encounter_num</u>	INTEGER
PK	<u>patient_num</u>	INTEGER
	inout_cd	VARCHAR(10)
	location_cd	VARCHAR(100)
	location_path	VARCHAR(700)
	start_date	DATETIME
	end_date	DATETIME
	visit_blob	TEXT(10)

M

E patient_dimension		
PK	<u>patient_num</u>	INTEGER
	vital_status_cd	VARCHAR(10)
	birth_date	DATETIME
	death_date	DATETIME
	sex_cd	CHAR(10)
	age_in_years_num	INTEGER
	language_cd	VARCHAR(100)
	race_cd	VARCHAR(100)
	marital_status_cd	VARCHAR(100)
	religion_cd	VARCHAR(100)
	zip_cd	VARCHAR(20)
	statecityzip_path	VARCHAR(200)
	patient_blob	TEXT(10)

CORE

A observation_fact		
PK	<u>encounter_num</u>	INTEGER
PK	<u>concept_cd</u>	VARCHAR(20)
PK	<u>provider_id</u>	VARCHAR(20)
PK	<u>start_date</u>	DATETIME
PK	<u>modifier_cd</u>	CHAR(1)
	patient_num	INTEGER
	valtype_cd	CHAR(1)
	tval_char	VARCHAR(50)
	nval_num	DECIMAL(10,2)
	valueflag_cd	CHAR(1)
	quantity_num	DECIMAL(10,2)
	units_cd	VARCHAR(100)
	end_Date	DATETIME
	location_cd	TEXT(100)
	confidence_num	VARCHAR(100)
	observation_blob	TEXT(10)

V
M

concept_dimension		
PK	<u>concept_path</u>	VARCHAR(700)
	concept_cd	VARCHAR(20)
	name_char	VARCHAR(2000)
	concept_blob	TEXT(10)

M

provider_dimension		
PK	<u>provider_path</u>	VARCHAR(800)
	provider_id	VARCHAR(20)
	name_char	VARCHAR(2000)
	provider_blob	TEXT(10)

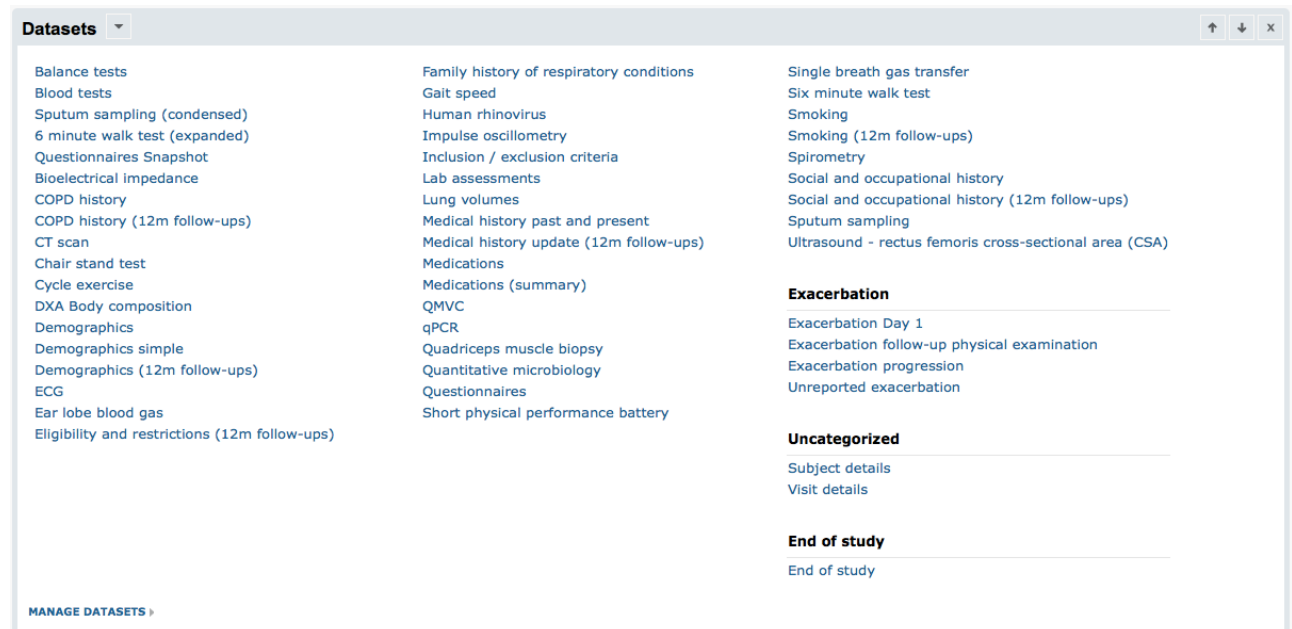
M



- £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



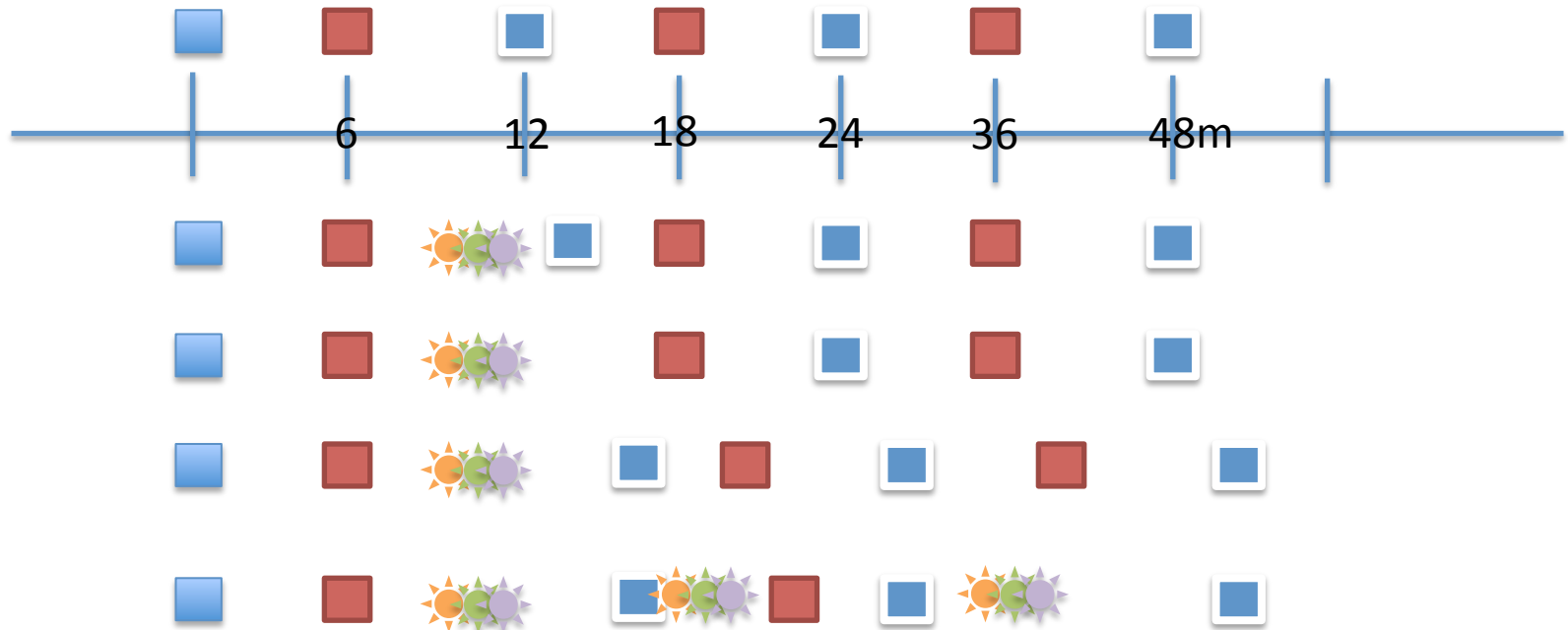
Exacerbation, plus follow-ups

525 patients

Up to 6 stable visits *plus*

0-9 exacerbation visits *plus*

2w + 6w follow-ups

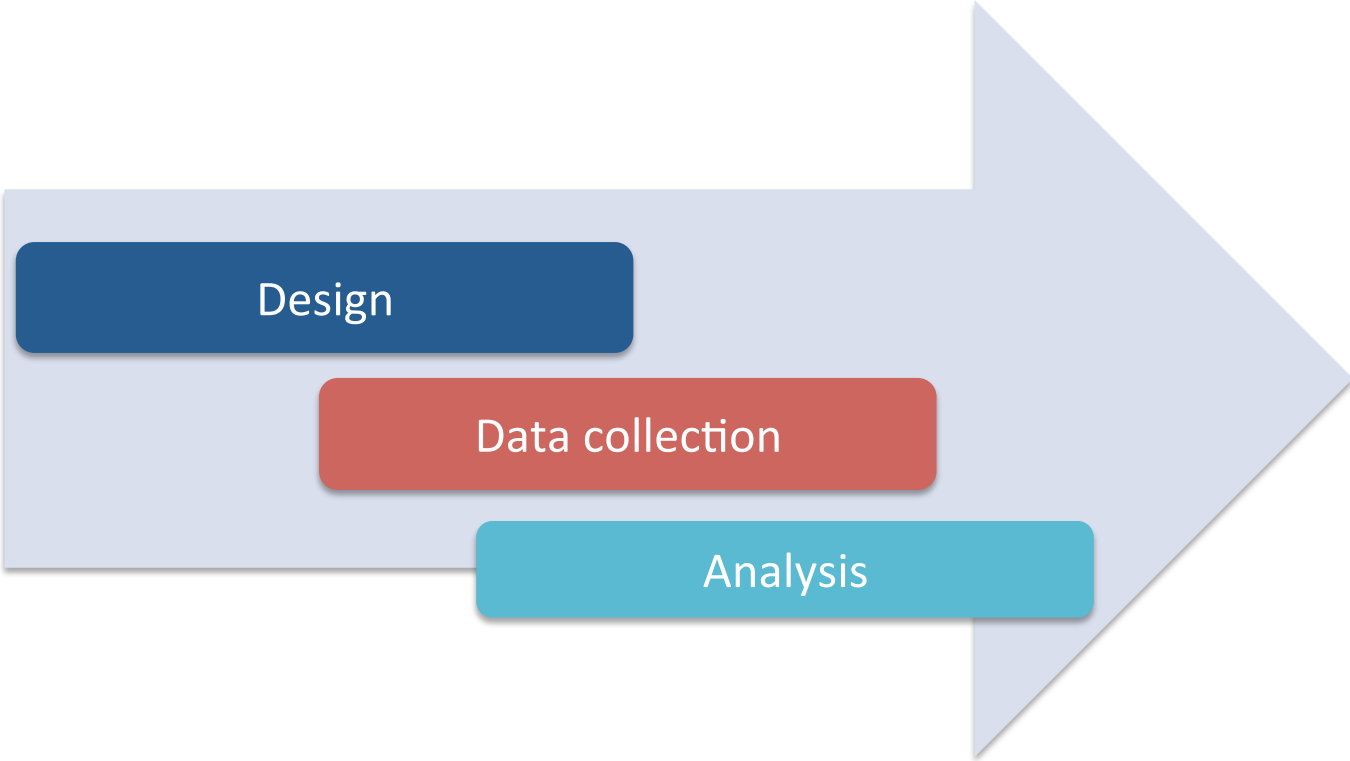




Design

Data collection

Analysis

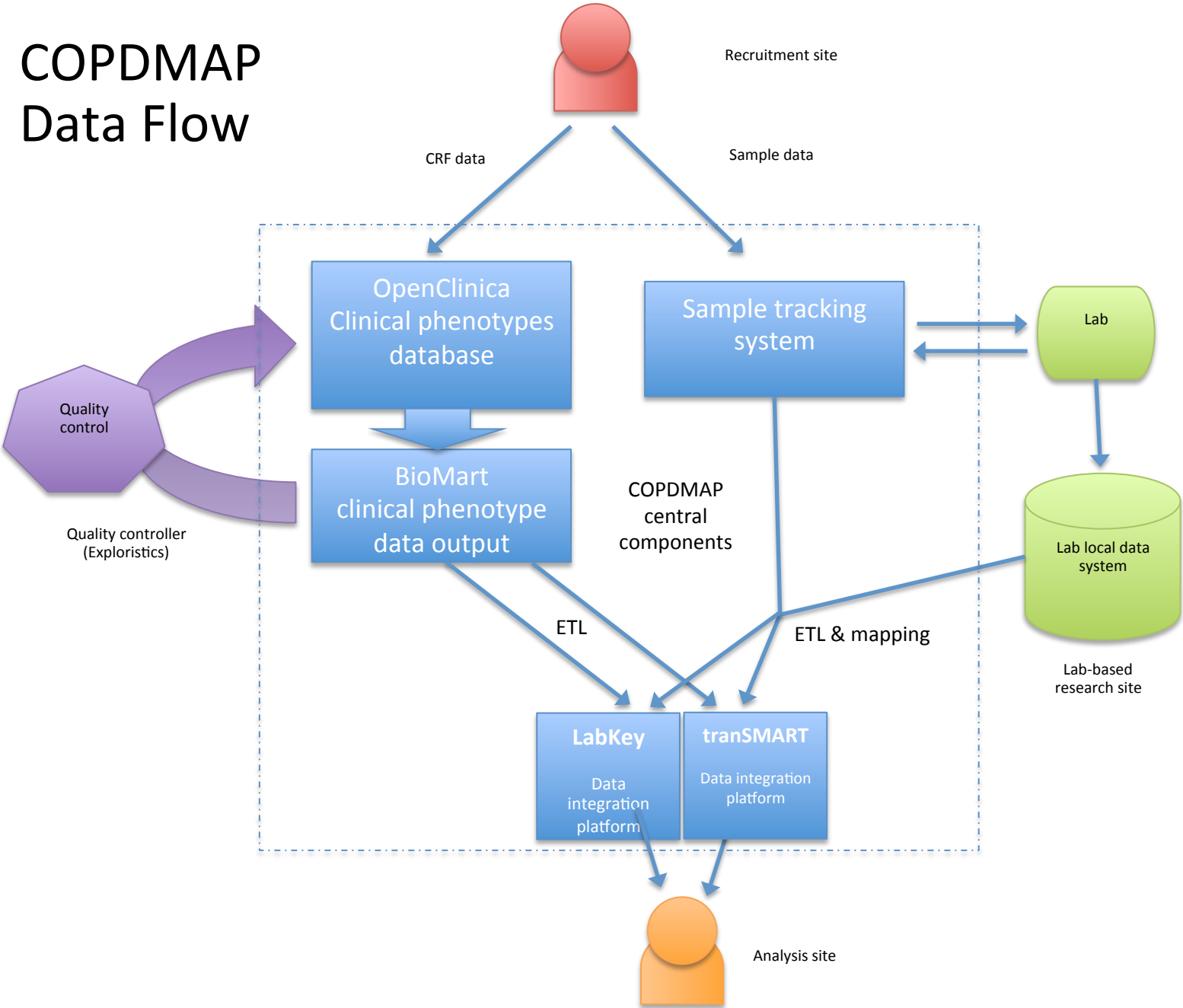


Design

Data collection

Analysis

COPDMAP Data Flow



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 COPDMAP data with other studies

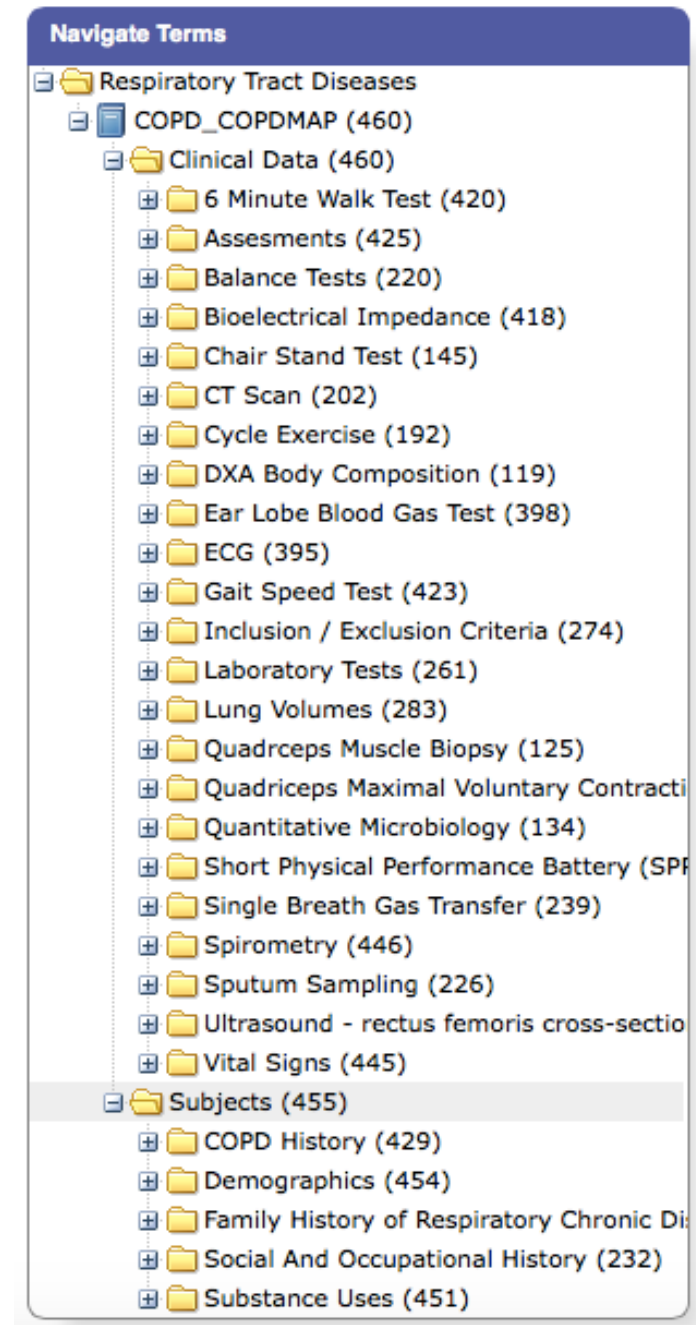
 - Workflow based approach.
 - Define cohort -> Summary data -> Analyse / Export
- Problems?
 - Lengthy ETL process
 - Limited fine-grained access control (by ontology nodes, not by subject)
 - Complex visit structure
 - Some data types not supported (e.g. full text, detailed medications, dates)

COPD MAP tranSMART tree

ETL:

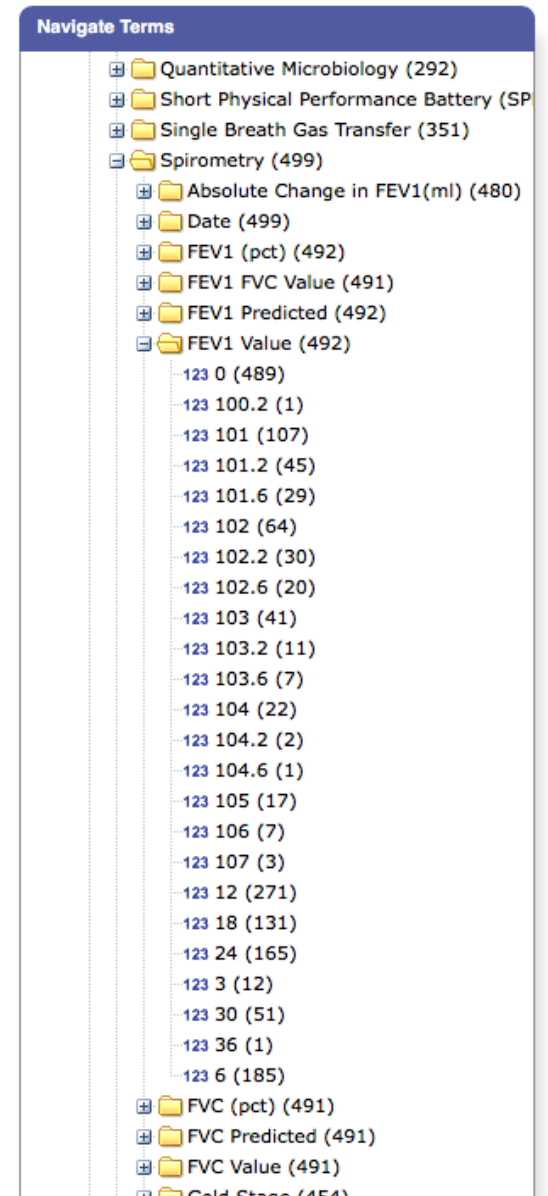
Baseline: Pfizer

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



UNIVERSITY OF
LEICESTER

Leicester Biomedical Research Centre: Cardiovascular Theme

Informatics team

Richard Bramley

Daniel Lawday

Susan Sterland

Shajid Issa

BioResource manager

Dr Gavin Whyman



***National Institute for
Health Research***