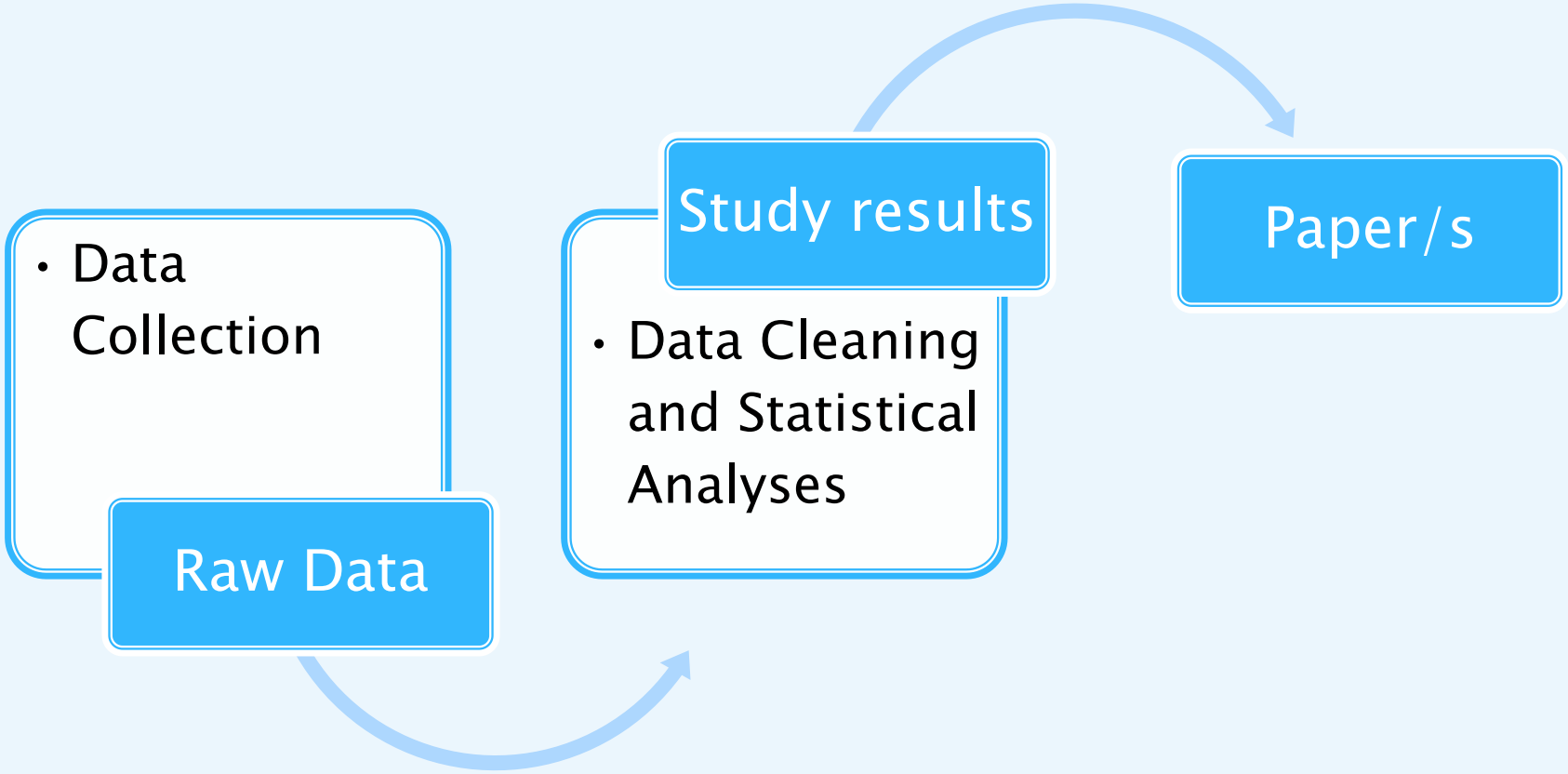
Big Clinical Data Management Automation

Antonella Delmestri, PhD
Senior Database Manager, CSM, University of Oxford

Traditional Clinical Data Pipeline



Big Clinical Data

- The amount of clinical data digitally collected and stored is vast and expanding rapidly
- Data cleaning accounts for 30%-80% of the development time and budget
- Any redundant or inconsistent data have the potential to confound analysis that aggregates or reasons from the data. It is essential to understand the extent and kind of problem, and to have methods for managing it.
  - Data Quality: Theory and Practice (W. Fan, 2012)

Pharmaco and Device Epidemiology group Approach

- Separate big data management and curation from statistical analyses
- Automate big data management and curation
- Automate variables extraction

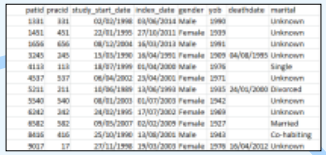
Big Clinical Data Pipeline

- Big Data Collection

Raw data (~Gb)

Variables (~Mb)

- Automated Big Data Management, Curation, Variables Extraction



- Statistical Analyses

Study results

Paper/s

Why splitting these two processes?

- To allow statisticians to focus on statistical analyses and innovation, e.g. prediction models and machine learning techniques
- Address statistical packages limitations to handle the multi-dimensionality and volume of big clinical data
- To benefit from Computer Science technologies to manage big clinical data

Automation – Why?

- To ensure research is:
  - Reproducible
    - Others may verify our findings and build upon them
  - Consistent
    - Over time, across items and across different studies
  - Reliable
    - High quality: accurate and precise
  - Useful
    - Protect public health
  - Fast
    - Efficient, competitive

Automation – How?

- Use and develop specialised software for Data Management, Curation and Extraction:
  - DataBase Management Systems (DBMS) e.g. MySQL
  - Programs written in a programming language suitable for the purpose (e.g. Python)

- Develop Standard Operating Procedures (SOPs) for Data Management
  - Multidisciplinary effort
  - Include them in the software

SOPs Example

- Exclude records that occur at unacceptable dates
  - E.g. After patient death
  - After date when data were downloaded
  - After last practice upload date
  - After transfer out of the practice date
- Exclude records reporting duplicated/inconsistent information for the same patient on the same date of event
  - E.g. patid = 1234, BMI = 25.8, date = 20/12/2016
  - patid = 1234, BMI = 28.5, date = 20/12/2016
- Exclude unacceptable values
  - E.g. BMI: <10 or >100

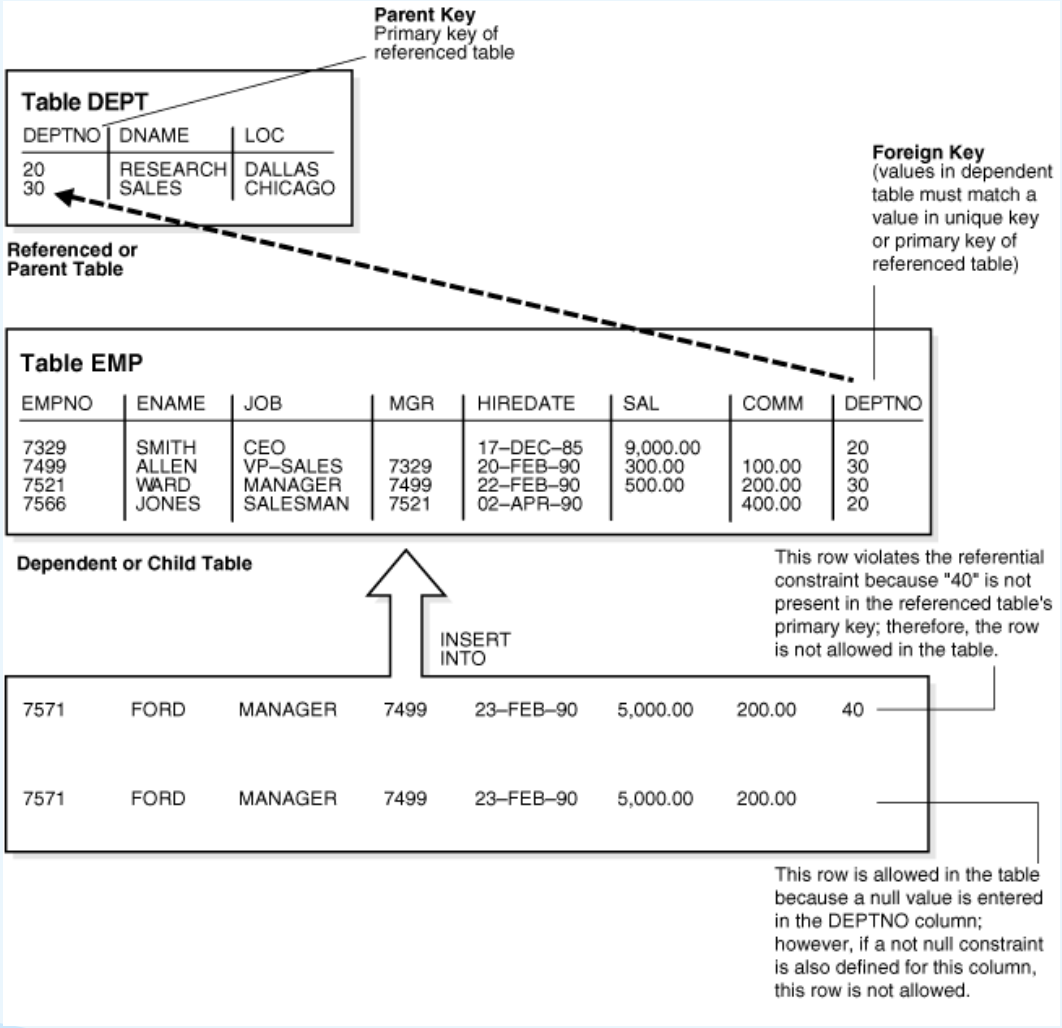
DataBase Management Systems (DBMSs)

- Database = **O**rganized (logically structured) collection of data as opposed to a "Dataset", which is just a collection of data
- DBMS = computer software application that interacts with users, other applications and the database itself. Golden standard for any data management
- Relational DBMSs (RDBMSs) since the 1980s support the relational model represented by SQL, the Standard Query Language (e.g. MySQL, Microsoft SQL Server, Oracle, Microsoft Access, etc.)
- RDBMSs benefits (selected):
  - No data redundancy
  - No data inconsistency

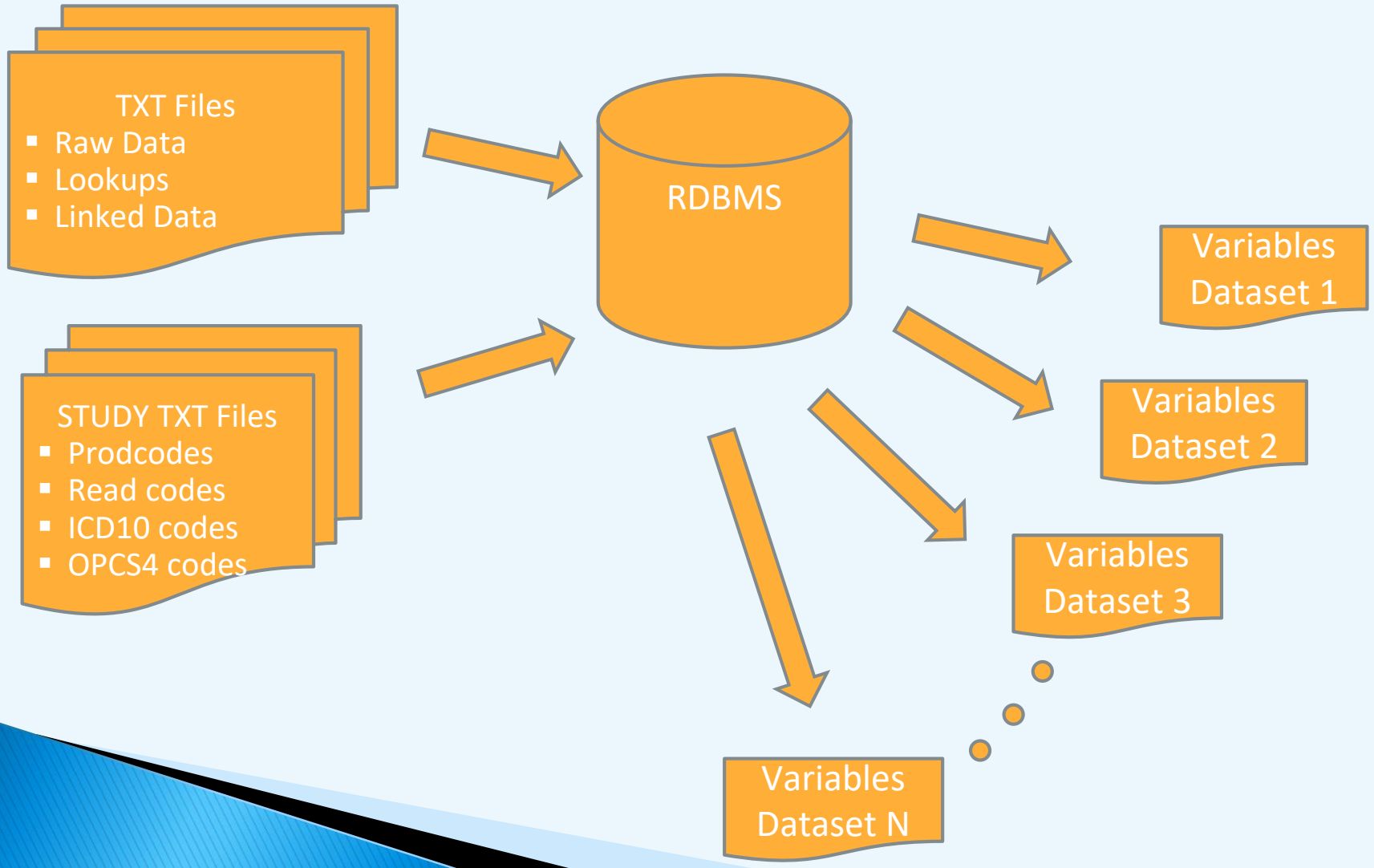
RDBMSs

- Primary Keys / Unique Keys
  - Entity integrity
- Data Types
  - Domain integrity
- Foreign Keys
  - Referential integrity
- Efficiency
  - Indexes
- Data Security
  - User access control
  - Data encryption

Referential Integrity



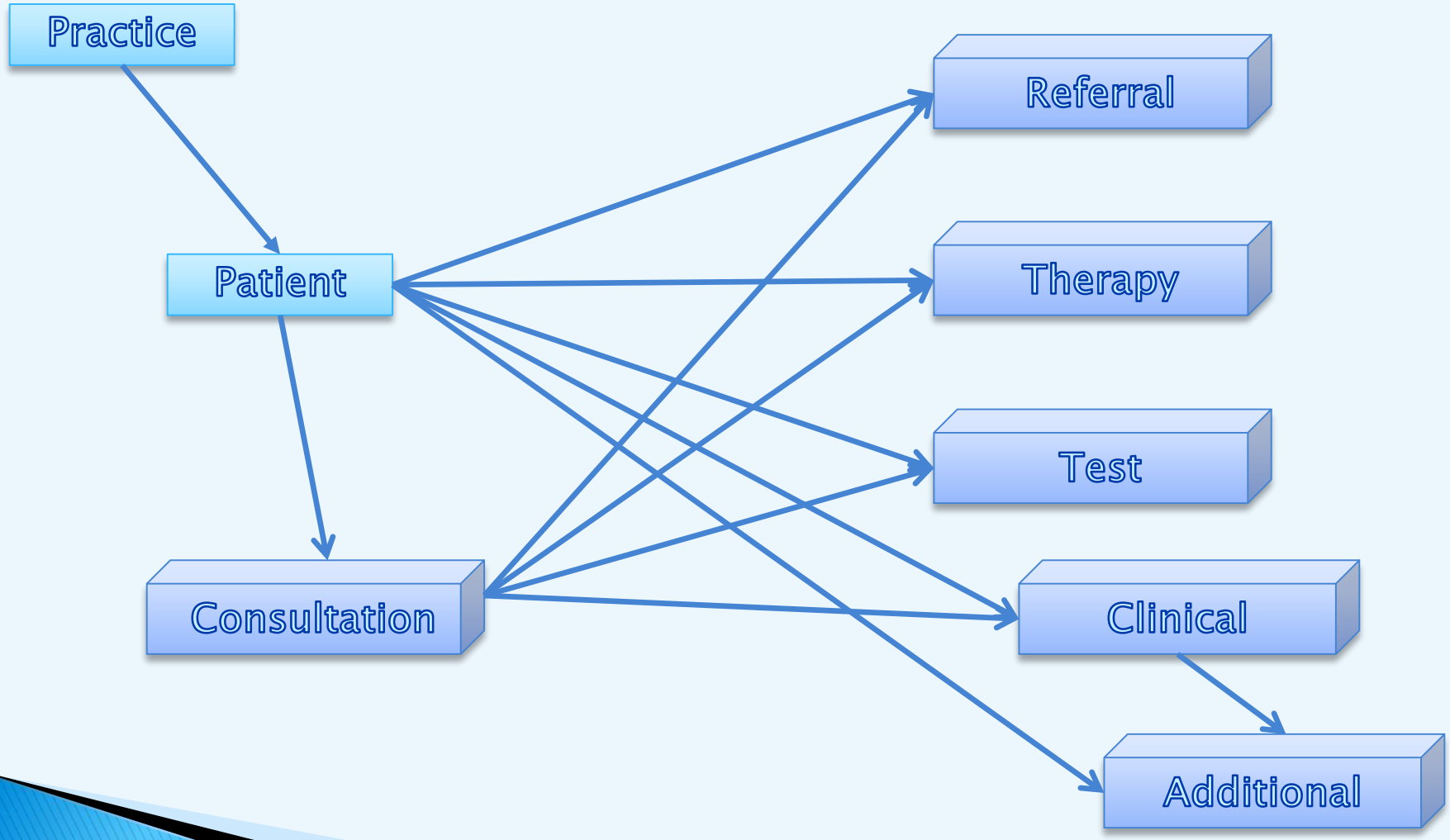
Data Flow


TXT Files
Raw Data
Lookups
Linked Data
STUDY TXT Files
Prodcodes
Read codes
ICD10 codes
OPCS4 codes
RDBMS
Variables Dataset 1
Variables Dataset 2
Variables Dataset 3
Variables Dataset N

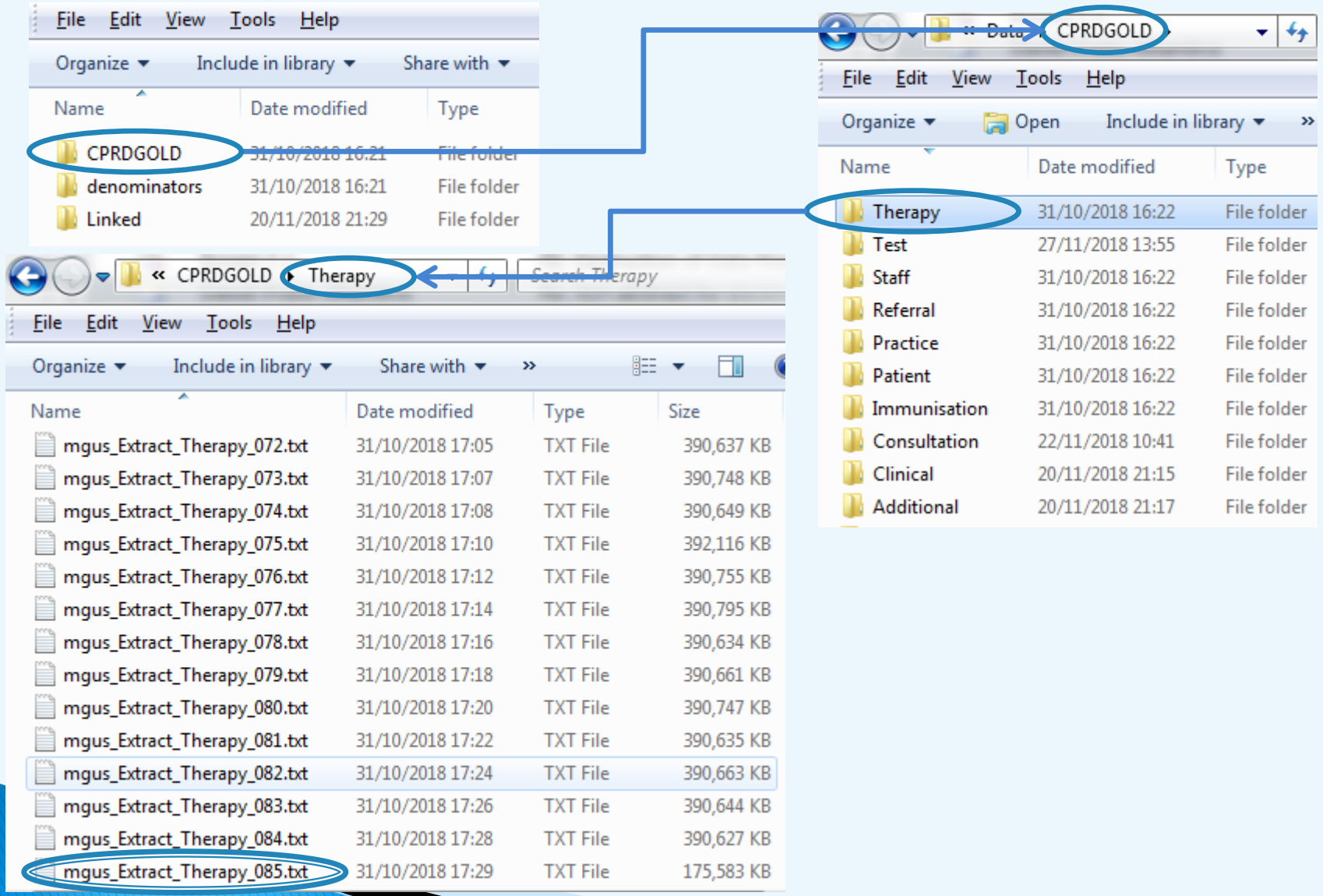
Automation – Phase 1 Database Design

- Understand dataset structure
  - Tables
  - Lookups
- Identify problems in the dataset structure, if any
  - Define sensible solutions
- Identify problems in raw data, if any
  - Define sensible solutions
- Write code
  - E.g. In python with embedded MySQL

CPRD GOLD Dataset Structure



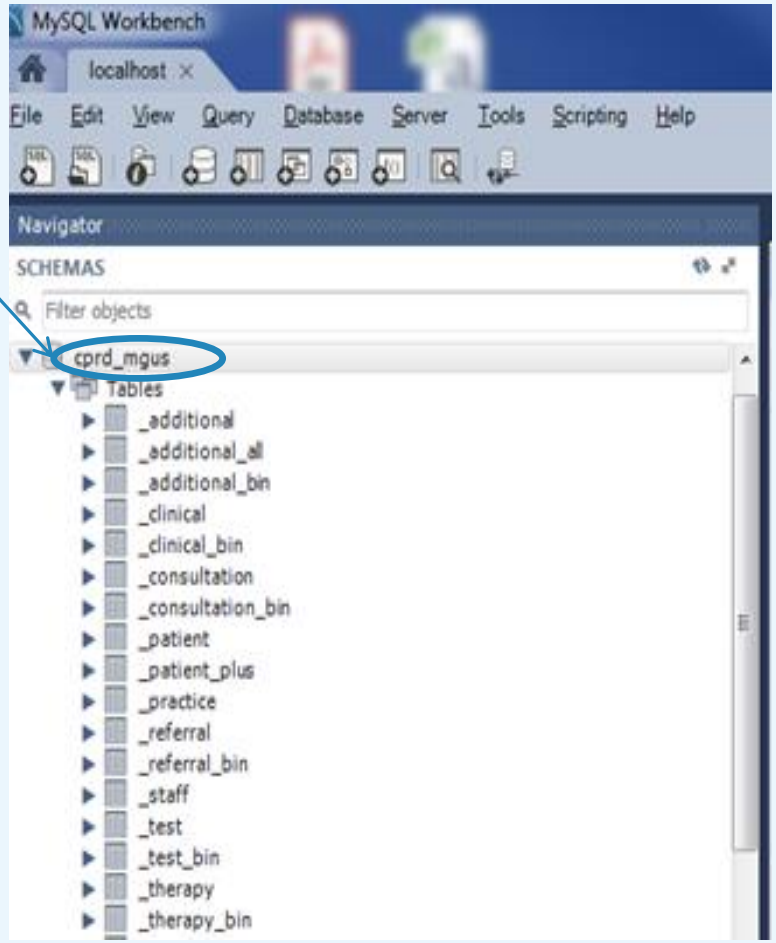
CPRD GOLD Raw Data



Automation – Phase 1 Reading Raw Data

- Configure database settings:
  - DATABASE_NAME = 'cprd_mgus'
  - GOLD_VERSION = 2.0
  - LINKAGE_SET =16
  - Etc.

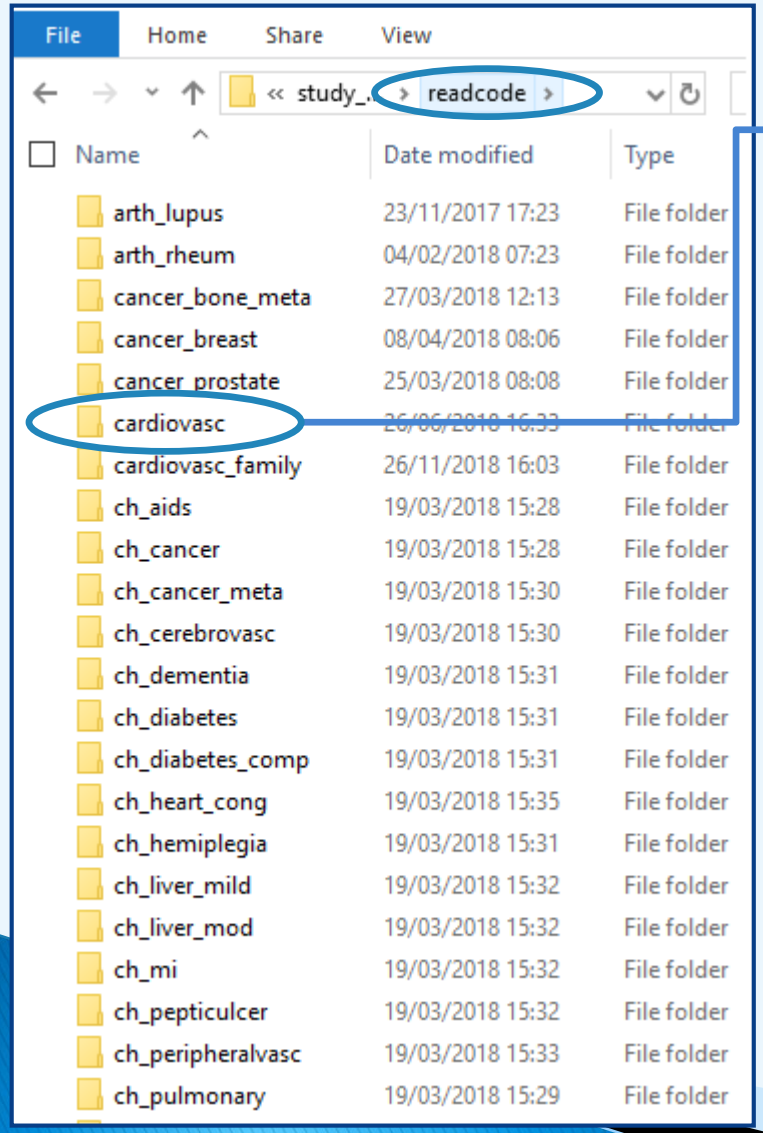
- Read raw data
  - Run code for database creation and population, which includes the bin tables where to put unacceptable records and the reason for it

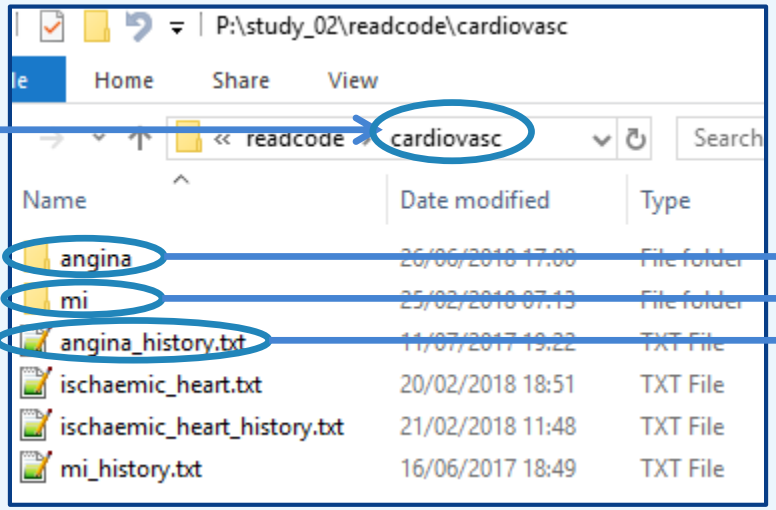


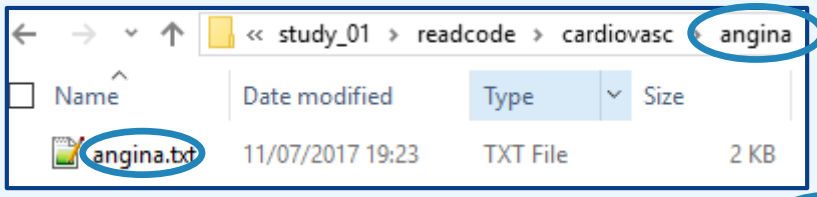
Variables Request – Clinical

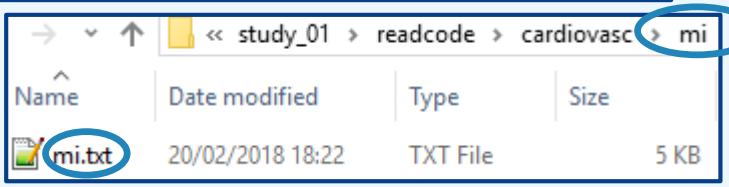
Name	Readcode/Medcode Files	Description1	Description2
LUPUS	arth_lupus.txt	Flag + Date of last diagnosis of systemic lupus erythematosus before or on index date	
RHEUMATOID ARTHRITIS	arth_rheum.txt, arth_rheum_flare.txt, arth_rheum_history.txt	Flag + Date of last diagnosis of rheumatoid arthritis before or on index date	
BONE METASTATIC	cancer_bone_metastasis.txt	Flag + Date of first ever (before, on or after index date) diagnosis of bone metastasis	
BREAST CANCER	cancer_breast.txt, cancer_breast_history.txt, cancer_breast_metastasis.txt	Flag + Date of first ever (before, on or after index date) diagnosis of breast cancer	
PROSTATE CANCER	cancer_prostate.txt, cancer_prostate_history.txt	Flag + Date of first ever (before, on or after index date) diagnosis of prostate cancer	
CARDIOVASC	angina.txt, angina_history.txt, ischemic_heart.txt, ischemic_heart_history.txt, mi.txt, mi_history	Flag + Date of last cardiovascular disease event before or on index date	
MI	mi.txt	Flag + date of MI event on index date.	Flag + Date of first MI event after index date
ANGINA	angina.txt	Flag + date of angina event on index date.	Flag + Date of first angina event after index date

Study Definition Taxonomy









- angina_history.txt
  - medcode 6336 57062
  - readcode 14A5.00 14AJ.00
  - readterm H/O: angina pectoris H/O: Angina in last year

Automation – Phase 1
Read Study Definition

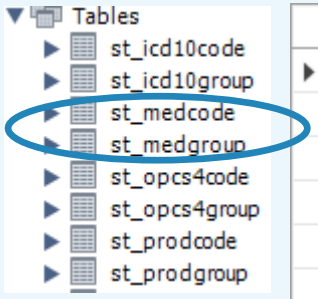


Table: st_medgroup

group_id	group_name	father_id
1	arth_lupus	NULL
2	arth_rheum	NULL
3	cancer_bone_meta	NULL
4	cancer_breast	NULL
5	cancer_prostate	NULL
6	cardiovasc	NULL
7	angina	6
8	mi	6
9	ch_aids	NULL
10	ch_cancer	NULL
11	ch_cancer_meta	NULL
12	ch_cerebrovasc	NULL
13	ch_dementia	NULL
14	ch_diabetes	NULL
15	ch_diabetes_comp	NULL
16	ch_heart_cong	NULL
17	ch_hemiplegia	NULL

Table: st_medcode
Group_id = 6 (including children: 7, 8)

readcode	readterm	group_id	medcode
14A3.00	H/O: myocardial infarct <60	6	35674
14A4.00	H/O: myocardial infarct >60	6	40399
14A5.00	H/O: angina pectoris	6	6336
14AH.00	H/O: Myocardial infarction in la...	6	50372
14AJ.00	H/O: Angina in last year	6	57062
14AL.00	H/O: Treatment for ischaemic h..	6	45476
14AT.00	History of myocardial infarction	6	100139
323..00	ECG: myocardial infarction	8	7783
3232.00	ECG: old myocardial infarction	6	39904
3236.00	ECG: lateral infarction	6	52705
323Z.00	ECG: myocardial infarct NOS	8	59032
662K100	Angina control - poor	7	15373
662K200	Angina control - improving	7	14782
662K300	Angina control - worsening	7	29300
7929100	Percut transluminal coronary th...	6	33650
7929111	Percut translum coronary thro...	6	40996
889A.00	Diab mellit insulin-glucose infus..	8	61670
G3...00	Ischaemic heart disease	6	240
G3...13	IHD - Ischaemic heart disease	6	1792
G30..00	Acute myocardial infarction	8	241
G30..11	Attack - heart	8	13566
G30..12	Coronary thrombosis	6	2491
G30..13	Cardiac rupture following myoc...	8	30421
G30..14	Heart attack	8	1204
G30..15	MI - acute myocardial infarction	8	1677

Automation – Phase 2 Variable Extraction

- Write library of Procedures with Parameters: e.g.

- get_clinical_event("mi", "on", index_date);
- get_clinical_event("mi", "after", index_date);

- get_prescription("calcium", "after", index_date);
- get_prescription("statin", "prior_or_on", start_date);

- get_icd10_event("icd10_fx", "prior", index_date);

- get_opcs4_event("opcs4_bariatric", "after", start_date);

Thank you

Questions