Importance of Data Quality in Cohorting

Dr Dai Evans

GP & Clinical Advisor, PRIMIS, University of Nottingham

COVID Vaccination example

- Female aged 15 on CHC (maternal knowledge) requests repeat Rx of CHC; process at time involves remote template information transfer including Weight in Kg
- Appropriate weight recorded -> BMI calculated by admin staff

- Last recorded height was 85cm from some years earlier -> BMI = 63
- This qualifies her for an “At Risk” group for COVID -> needs immunisation

- Called attends and correctly [!] given Pfizer vaccine as age 15
- Vaccination incorrectly recorded as AZ on POC system

- Recalled and as AZ recorded on system – given AZ dose for 2nd jab …..

How do we use Cohorting in the NHS?

- To identify groups of patients:
  - Who may benefit from direct patient care – COVID vaccination
  - Where their care could be improved – diabetes, QoF, PINCER
  - To help planning and resource allocation – workload, deprivation
  - To assist research

Data Quality

- DQ depends on the use you are going to put that data to

  eg medical record information as an aide-memoire

- So usually, Cohorting is *not* the primary use of the recorded information and therefore the required data may be missing

  eg ethnicity data in COVID or deprivation studies

  eg gaps in the record of immune suppressive treatments delivered in Secondary Care

Types of “poor” DQ (from cohorting perspective)

- Missing data
  - complete missing item *eg MI, specific medication, recent BP*
  - dependant item missing/outdated -> erroneous data *eg old height in BMI*
  - missing items in wider dataset – may not actually impact
    eg in algorithm can cope with 1-2 missing variables out of 10
- Wrong data
  - wrong code/term or even in wrong record
  - wrong contact details
  - wrong date
  - association of inappropriate text altering meaning of term
- Imprecise data item *eg CVA without detail of whether haemorrhagic or not*
- Effects of Legacy Artefacts *eg vaccination data*

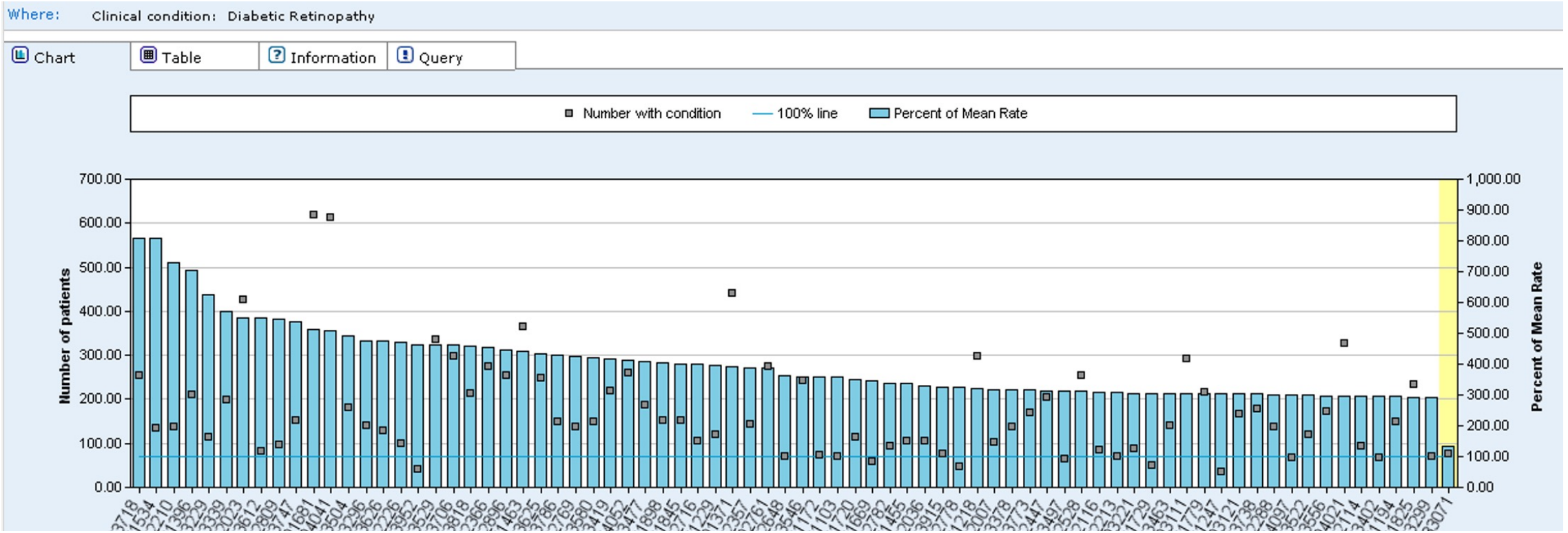
FLU vaccination data

- UKHSA monitors various aspects of vaccination programmes
- Includes efficacy of vaccine *type (need data on actual vaccine used.*)*

- Monitor recording of “Seasonal Influenza Vaccination” codes
  - Alerted to persistence of *Pandemic Influenza Vaccination* codes use
  - Detailed analysis showed inter alia hinting at persistence of legacy artefacts**:

SNOMED_Concept_ID	Description	Usage
515281000000108	PANDEMRIX - first influenza A (H1N1v) 2009 vaccination given (procedure)	68,070
515301000000109	PANDEMRIX - second influenza A (H1N1v) 2009 vaccination given (procedure)	2,460
515291000000105	CELVAPAN - first influenza A (H1N1v) 2009 vaccination given (procedure)	440
515321000000100	CELVAPAN - second influenza A (H1N1v) 2009 vaccination given (procedure)	140

Template content error: *Retinopathy*



Research example: incomplete dataset

- FLUCAT project
- Identified core dataset from previous epidemic/pandemics
- Additional data items included from current experience

- Some recorded data items not in “available” extracted datasets GPDPR

  eg oxygen saturation

Another Type of Poor DQ: *Data Handling Errors*

- Data set definition may be flawed – **wrong codes included/excluded/missing ***
- Time frames or other logical rules may be flawed
- Erroneous data cleansing (steps in analytical preparation)
- Being unaware of implications of Terminology shifts, tooling or related dependencies ***eg V2 -> CTV3 -> SCT; hierarchical shifts***

	DJE's % of above	10.7%

Consultation Problem Titles added
(Top 5 areas are highlighted)

Area	No. of codes	% of Total
Infectious and parasitic diseases	20	1.2%
Neoplasms	34	2.0%
Endocrine	263	15.3%
Diseases of Blood	23	1.3%
Mental disorders	99	5.8%
Nervous system diseases	74	4.3%
Circulatory system diseases	162	9.4%
Respiratory system diseases	67	3.9%
Digestive system diseases	34	2.0%
Genitourinary system diseases	37	2.2%
Pregnancy	4	0.2%
Skin diseases	91	5.3%
Musculoskeletal diseases	156	9.1%
Congenital anomalies	0	0.0%
Perinatal conditions	0	0.0%
Ill defined conditions	77	4.5%
Injury & Poisoning	28	1.6%
Causes of injury and poisoning	7	0.4%
Morbidity and mortality	0	0.0%
Unspecified conditions	7	0.4%
History/Symptoms	304	17.7%
Examinations/Signs	7	0.4%
Diagnostic procedures	6	0.3%
Laboratory procedures	10	0.6%
Radiology	0	0.0%
Preventative procedures	29	1.7%
Operations	13	0.8%
Other therapeutic procedures	14	0.8%
Administration	24	1.4%
Other	121	7.1%
Unassigned	5	0.3%
Total	**1,716**	**100.0%**

3 month Workload Analysis

- Depends on coding problem titles in consultations
- Also depends on correct data analysis

- *Permits organisational questions*

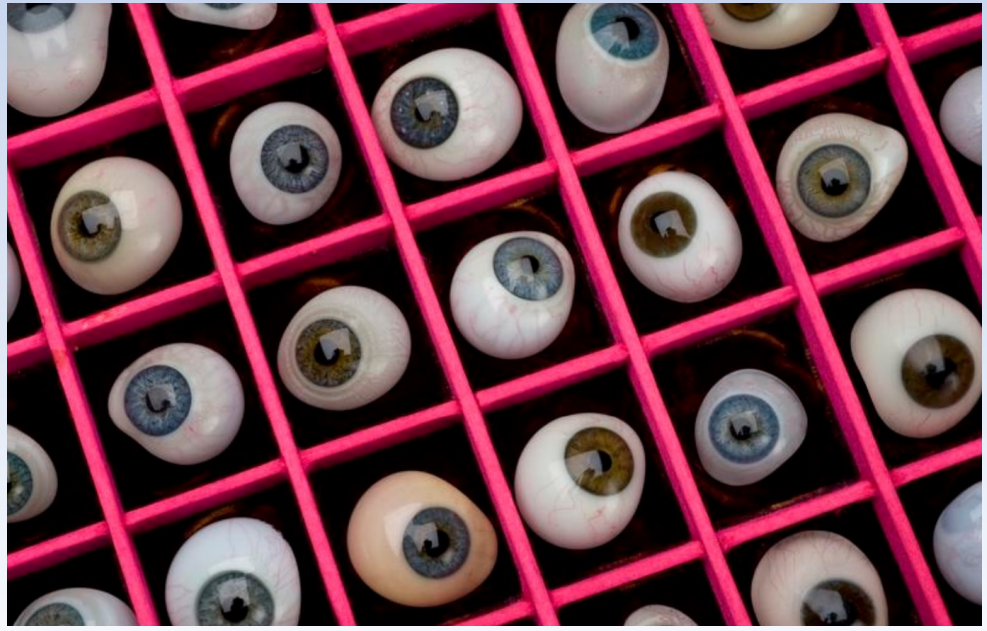
Staff Category	Consultations	-	-	-	-	-	-	5
Physio		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

- *Leads to organisational solutions*

How to get it right?

- Data Entry education and training…. ***All those entering clinical data***
- Expose the data/information to patients
- Care with Terminologies ***(especially evolving SCT, note shifts in usage)***
- Review of “legacy” artefacts
- Use of correct data entry artefacts – **templates, protocols etc**
- Inbuilt validation rules ***eg BMI checker***
- Check data extraction routines
  - ***second pair of eyes,***
  - ***compare/use other sources***

- **Sense check your results!**



Other solutions

- Some concepts are not included in current terminology usage
- Some concepts are not included in current terminology.
- Consider requesting release of codes outside GP subset
- Consider requesting new codes to cover gaps in terminology
- Check coding frequencies – *NHS-D code frequency tables*
- Consider other data sources – *HES, National Datasets etc*
- Consider data validation checks *eg qualifying HbA1c for T2DM*
- DQ Feedback to bespoke contributing networks

“Cohorting” process

- Define your question
  - What are you trying to do with the data?
  - Precise unambiguous question (& component elements) is required
  - Is it feasible? Are there things you cant get or are unlikely to get? Do you have to think laterally to answer your question?
- Define both your search criteria AND your output data
- Test run -> sense check of the results & does it answer your need
- Be prepared to amend it (*unforeseen problems*)
- Remember if repeating over time – *may need to amend* due to terminology shifts

Questions and discussions

Dressing for the
dinner at PHCSG