cdisc

2023
EUROPE
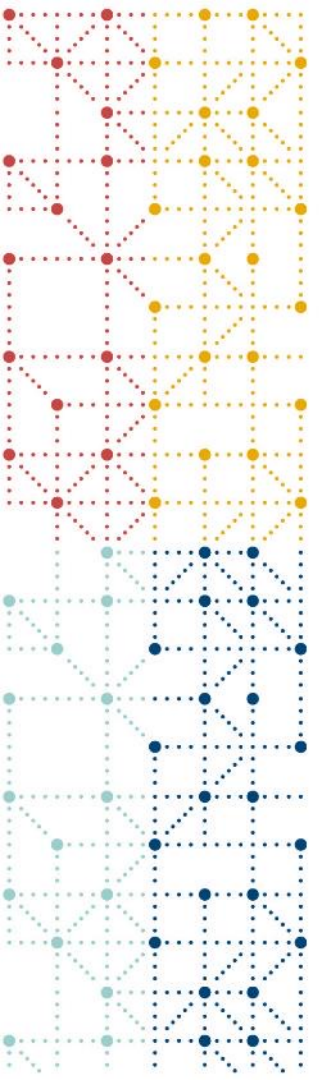## INTERCHANGE
COPENHAGEN | 26-27 APRIL

Creation of SDTM Datasets for a Real World Study using Electronic Medical Records

Presented by Norman Stein, Senior Statistician, Data Science and Analytics Department, NWEH

Meet the Speaker

Norman Stein

Title: Senior Statistician

Organization: NWEH

Norman Stein studied mathematics at Cambridge University before working for several years in the engineering industry. After taking a PhD in theoretical physics at Manchester University, he worked in a variety of positions in academia and industry, on topics including noise-driven nonlinear systems and protein crystallography. For the past thirteen years he has been a statistician and data scientist at NorthWest EHealth in Manchester, working on real world clinical trials such as the Salford Lung Study, as well as retrospective data studies using electronic medical records.

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

- *The author(s) have no real or apparent conflicts of interest to report*

Agenda

1. Background - UK Healthcare system
2. eCRF and EMR based SDTM tables
3. Populating and Coding the MH and CM tables
4. Secondary Care Data

Background – the UK Healthcare System

Introduction to the UK Healthcare System
Primary and Secondary Care

Divided into primary care (GP/family doctors) and secondary care (hospitals)

Primary care traditionally coded using V2 Read codes

Secondary care uses:

- ICD10 codes for diagnoses
- OPCS4 codes for operations
- HRG codes for costs

Change to SNOMED – CT is slowly occurring

ICD10 codes are used internationally

OPCS4, HRG and Read codes are UK specific

cdisc

Real World Trials

Conducted in a primary (or secondary) care setting

Number of trial-specific visits kept to a minimum

Participants continue to receive usual care from GPs and hospitals

Data collected electronically from EMR feeds rather than eCRF

Measure efficiency rather than efficacy

This presentation describes work done on an actual clinical trial

Interventional, prospective study with approximately 1000 participants

cdisc
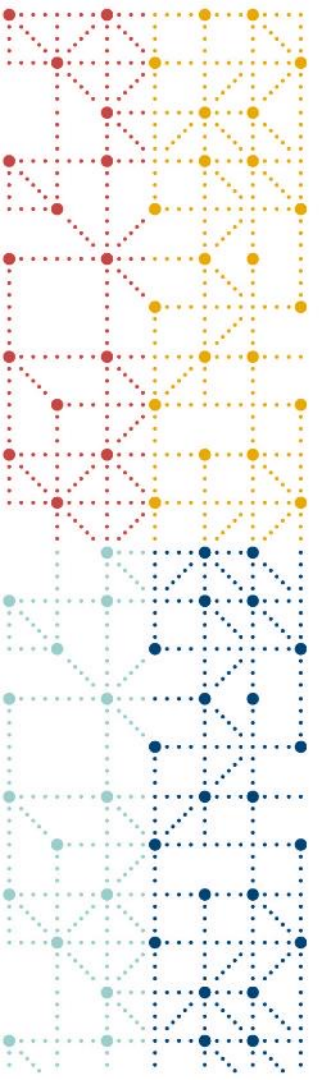
eCRF and EMR based SDTM tables

SDTM dataset creation – eCRF based data

SDTM datasets created for clinical trial with big pharma company

Majority of SDTM tables use eCRF data

Data entered into eCRF by study team (research nurses, PIs, and delegates).

Collected into a database using NWEH proprietary software

Stata scripts:

- access data using ODBC
- manipulate as necessary
- write output data to SDTM tables

Process is generally straightforward

cdisc

Reasons for using an EMR

Some domains, e.g. CM and MH lend themselves to EMR

Sheer volume of data makes it difficult for research nurse to enter data into an eCRF in a reasonable timescale

EMR reduces likelihood of transcription errors

Other domains, e.g. LB may use EMR as a source of additional data

Other advantages include ability to perform real time safety monitoring
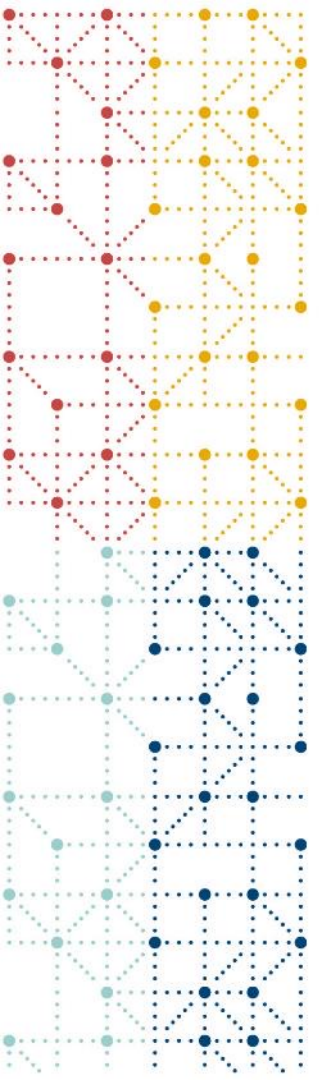
Populating and Coding the MH and CM Domains

Creation of the Medical History (MH) table

Diagnoses have Read codes beginning with letters A-T

Diagnoses also have a Read Term (MHTERM)

Data from all GP practices on study extracted by a third party

Diagnoses extracted from primary care record

May want to retain only the earliest example of a particular diagnosis

MedDRA coding uses a variety of techniques:

- Exact match between Read Term and MedDRA lowest level term (llt).
- Read code → SNOMED code → MedDRA llt
- Manual look up
- Automatic algorithm

cdisc

MedDRA coding results

Method	Number of matches	Percentage
Exact match	1464	42.1%
Read code → SNOMED code → MedDRA llt	790	22.7%
Automatic algorithm	624	17.9%
Manual look-up	602	17.3%

Creation of the Concomitant Medications (CM) table

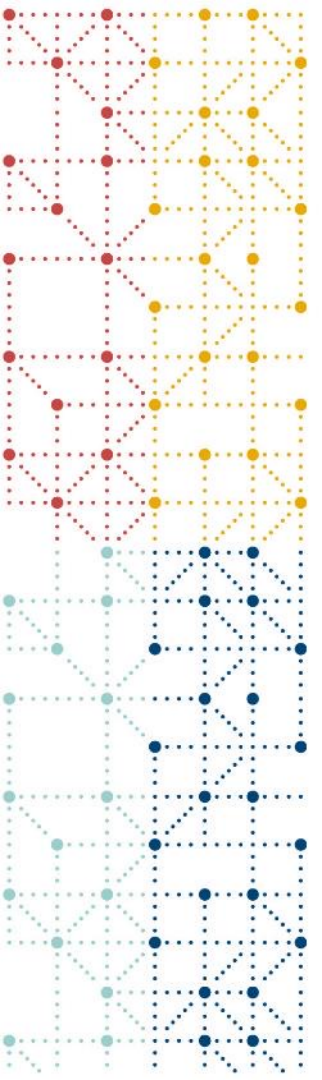EMR contains details of each prescription issued to patient

Structure in CM:

- One record per prescription
- CMSTDTC = prescription issue date
- CMENDTC is blank

A few variables included in SUPPCM:

- Number of tablets/capsules supplied
- Cost (from Drug Tariff)
- Dosage directions – free text field – did not match standard codelist well

ATC (WHODRUG) coding straightforward

Secondary Care Data

Structure of Secondary Care Data

Data supplied by a national supplier as separate tables

- Inpatient admissions (planned/emergency, day-case/longer stays)
- Accident and Emergency
- Outpatients

Tables are very wide – there can be many diagnosis and operation codes

Costing is done based on several factors:

- Operations performed
- Diagnoses made
- Medical Specialty
- Length of Stay

cdisc

Creation of pseudo-SDTM tables for Secondary Care

Data did not fit well into standard format

A column-based format was adopted, similar to input data

Standard variables such as STUDYID and USUBJID were added

More use of derived variables/manipulation than customary

- Deduplication and pre-sorting was required

- Data may not be received in expected form
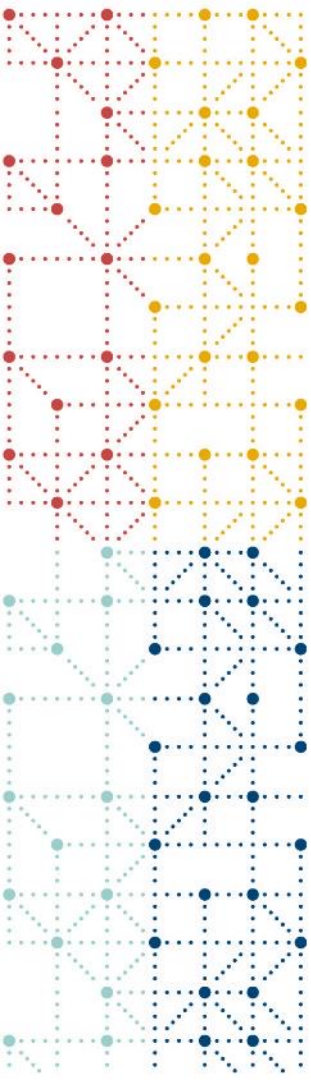
- Missed appointments should be removed

Data used as a source of surgical history data and MedDRA coded

cdisc

Other advantages of EMR data

Can use for systematic detection of Protocol Deviations

- Forbidden medications
- Study visits outside window

cdisc

Thank You!

This work was performed in collaboration with:

Sylwia Migas

John McCrae

NWEH

cdisc