SWAT 158: Validation of clinical events in the DaRe2THINK data-enabled clinical trial

Objective of this SWAT
Objective: To cross-validate coded primary healthcare data and secondary healthcare data in the DaRe2THINK data-enabled randomised trial using information extracted from a NHS Trust electronic healthcare record (EHR) database; for cardiovascular mortality, cerebrovascular events, thromboembolic events, myocardial infarction, vascular dementia, intracranial bleeding and other clinically-relevant bleeding events, where these events occur locally.

Aims:

1. Determine the accuracy of primary care outcomes derived from Clinical Practice Research Datalink (CPRD) database; percentage of patients with correct coding compared to Hospital Episode Statistics (HES) and EHR.

2. Determine the accuracy of secondary care outcomes derived from HES; percentage of patients with correct coding using a) primary codes, b) secondary codes, and c) both primary and secondary codes; compared to EHR.

3. Identify additional unreported events in the EHR; percentage of patients with clinical events of interest that are missing from either primary or secondary care datasets

Study area: Outcomes
Sample type: Patients
Estimated funding level needed: Medium

Background
DaRe2THINK (ISRCTN21157803, www.birmingham.ac.uk/d2t) is a pragmatic trial designed to answer a question with important public health impact: Does direct oral anticoagulant (DOAC) therapy reduce premature death, stroke and other thromboembolic consequences in younger patients with atrial fibrillation (AF), including prevention of cognitive decline and vascular dementia? The trial uses innovative methods to screen, recruit and follow-up patients, and is embedded within the UK National Health Service (NHS). Data on clinical events in the trial are entirely derived from the coded electronic healthcare records (EHR) across primary and secondary care using the Clinical Practice Research Datalink (CPRD) linked with national Hospital Episode Statistics (HES) and Office for National Statistics (ONS). The high accuracy of healthcare coding in the UK was demonstrated in a systematic review for hospital discharge codes used after the introduction of 'Payment by Results' in 2004.[1] Due to the General Medical Services 'Quality and Outcomes Framework' contract, accurate coding is also rewarded and incentivised in Primary Care. Multiple studies have validated disease-specific accuracy of CPRD data, including for complex diseases [2] and behavioural conditions.[3]

This SWAT will run alongside the main DaRe2THINK trial and assess the accuracy of the coded data from CPRD, HES and ONS in trial patients who attend the University Hospitals Birmingham (UHB) NHS Foundation Trust. Including four hospitals across Birmingham, UHB is one of the largest healthcare providers in Europe, treating more than 2.2 million patients each year. The large geographical footprint and provision of secondary, tertiary and quaternary care provides a unique ability to capture a wide variety of endpoints in patients recruited in the West Midlands. UHB has one of the most sophisticated electronic healthcare systems in the world, which has been in active use at for over 10 years. This includes the Patient Information Communication System (PICS), which integrates prescribing, drug administration, observation charting, diagnostic coding, laboratory and radiology tests. These data are used to enhance safe and efficient patient care through rules-based clinical decision support.

For patients recruited into DaRe2THINK, we will search our integrated healthcare records at UHB for pertinent data, for example an admission due to stroke or an outcome related to dementia. Searches will be based on NHS number, date of birth and name (recruitment to the trial includes informed consent for this purpose). We will systematically identify relevant clinical datasets

including healthcare records, clinical notes, imaging and time-series data. These structured, semi-structured and unstructured datasets will be aligned, harmonised and integrated across various modalities before application of machine learning behind the NHS firewall. Using Natural Language Processing (NLP) and text mining approaches, we will generate an automated, semantic characterisation of clinical endpoints for each patient. This will be compared with the outcome events obtained from linked EHR records to generate a statistical representation of the frequency and accuracy of outcomes.

Interventions and comparators
Intervention 1: Clinical outcome events of the host trial derived from CPRD data, coded data of primary care.
Intervention 2: Clinical outcome events of the host trial derived from HES data, coded data of secondary care, using primary codes, secondary codes or combination of primary and secondary codes.
Intervention 3: Clinical outcome events of the host trial derived from the combination of linked CPRD, HES and ONS coded data.
Intervention 4: Clinical outcome events of the host trial identified from the patient's EHR data at UHB NHS Trust.

Index Type: Other

Method for allocating to intervention or comparator
Other

Outcome measures
Primary: Percentage of patients with concordance of coded datasets for cardiovascular mortality, cerebrovascular events, thromboembolic events, myocardial infarction, vascular dementia, intracranial bleeding, and other clinically-relevant bleeding events.
Secondary: Additional unreported events; percentage of patients with clinical events of interest that are missing from either primary or secondary care datasets.

Analysis plans
Clinical outcome events will be identified in coded datasets (CPRD Aurum, HES and ONS) as well as the patient information extracted from local UHB data. We will use the clinical events identified from different sources to compute different measures of agreement, such as kappa statistic and inter-annotator agreement. This will allow us to highlight points of disagreement, at which point statistical analysis will be used to identify patterns of disagreement between sources and we will identify the optimal derivation of outcome variables and report on outcome agreement between resources. A Statistical Analysis Plan will be completed before any analysis commences and be made available on the DaRe2THINK website.

Possible problems in implementing this SWAT
Too few patients in the main DaRe2THINK trial attending the UHB NHS Trust.

References
1. Burns EM, Rigby E, Mamidanna R, et al. Systematic review of discharge coding accuracy. Journal of Public Health 2012;34:138-48.
2. Quint JK, Mullerova H, DiSantostefano RL, et al. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). BMJ Open 2014;4:e005540.
3. Hagberg KW, Jick SS. Validation of autism spectrum disorder diagnoses recorded in the Clinical Practice Research Datalink, 1990-2014. Clinical Epidemiology 2017;9:475-82.
4. Slater L, Bradlow W, Hoehndorf R, et al. Komenti: A semantic text mining framework [preprint]. bioRxiv 2020.08.04.233049.
5. Slater LT, Bradlow W, Ball S, et al. Improved characterisation of clinical text through ontology-based vocabulary. Journal of Biomedical Semantics 2021;12:7
6. Slater LT, Bradlow W, Motti DF, et al. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. Computers in Biology and Medicine 2021;130:104216.

Publications or presentations of this SWAT design

Examples of the implementation of this SWAT

People to show as the source of this idea: Xiaoxia Wang
Contact email address: Xiaoxia.Wang@uhb.nhs.uk
Date of idea: 10/MAY/2019
Revisions made by:
Date of revisions: