Logo

Description automatically generated

DATA ANALYSIS PLAN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

DPhil thesis:

A tale of three trials: exploration, validation, and application of routinely collected data on medications in large-scale randomised clinical trials

Version 0.2

Date: 15th November 2021

Author:

Guilherme Pessoa-Amorim

Supervisors:

Martin Landray

Louise Bowman

Will Stevens

Nuffield Department of Population Health

St Edmund Hall

University of Oxford

Contents

[1. Document history 3](#_Toc87965164)

[2. Abbreviations 4](#_Toc87965165)

[3. Introduction 5](#_Toc87965166)

[4. Background 5](#_Toc87965167)

[4.1. Rationale 5](#_Toc87965168)

[4.2. Objectives 5](#_Toc87965169)

[4.2.1. Primary 6](#_Toc87965170)

[4.2.2. Secondary 6](#_Toc87965171)

[4.2.3. Subsidiary 7](#_Toc87965174)

[4.3. Study design 8](#_Toc87965176)

[4.3.1. Data collection 8](#_Toc87965177)

[4.3.2. Overview of data sources 9](#_Toc87965178)

[4.3.3. Definition of study outcomes 11](#_Toc87965179)

[4.3.4. Main study hypothesis 17](#_Toc87965181)

[4.3.5. Analysis populations and sample size 17](#_Toc87965182)

[4.3.6. Definition of drugs of interest 18](#_Toc87965183)

[4.3.7. Derivation of drug groups and codelists 23](#_Toc87965184)

[4.3.8. Definition of reference standards 26](#_Toc87965185)

[4.3.9. Definition of time periods and time points 27](#_Toc87965186)

[5. Descriptive analyses 31](#_Toc87965187)

[5.1. Participant throughput 31](#_Toc87965190)

[5.2. Baseline characteristics of patient populations 31](#_Toc87965191)

[6. Comparative analyses 34](#_Toc87965192)

[6.1. Baseline characteristics 34](#_Toc87965193)

[6.2. Primary, secondary, and subsidiary outcomes 34](#_Toc87965194)

[6.3. Significance levels 35](#_Toc87965195)

[7. Subsidiary analyses 36](#_Toc87965196)

[7.1. RCD source features 36](#_Toc87965197)

[7.2. CDISC alignment 38](#_Toc87965199)

[7.3. Continuous measures of medication adherence 38](#_Toc87965200)

[7.4. Codelist development methods 39](#_Toc87965201)

[7.5. Other exploratory assessments 40](#_Toc87965202)

[8. Data monitoring 41](#_Toc87965203)

[9. Software 41](#_Toc87965204)

[10. References 43](#_Toc87965205)

[11. Tables 45](#_Toc87965206)

[12. Appendix 49](#_Toc87965207)

# Document history

|  |  |  |
| --- | --- | --- |
| Version | Date | Changes |
| 0.1 | 21-09-2021 | Initial draft |
| 0.2 | 15-01-2021 | 1. Revisions to text (WS) 2. Other changes (GPA):   Rearranged codelist compendium  Clarifications to section 4.2.2 and 6.1  Expanded section 6.2 and 7.5  New figure 1  Rearranged figure 2  New sections 4.2.3.4, 0, and 12 (Appendix) regarding codelists  Redefinition of event date derivation in the GP dataset (section 4.3.9.2)  Redefinition of time points and time periods for ORION-4 (section 4.3.9) |

GPA – Guilherme Pessoa-Amorim

WS – Will Stevens

# Abbreviations

ACEi: angiotensin-converting enzyme inhibitors

ARBs: angiotensin-receptor blockers

ARNIs: angiotensin-receptor/neprilysin inhibitors

BNF: British National Formulary

CCBs: calcium-channel blockers

CM: Concomitant Medication domain

CTSU: Clinical Trial Service Unit

eCRFs: electronic case report forms

GDPPR: General Practice Data for Pandemic Planning and Research

GLP1: Glucagon-Like Peptide 1

LDLc: low-density lipoprotein cholesterol

LMWHs: low-molecular weight heparins

MRAs: mineralocorticoid receptor antagonists

NHSBSA: National Health Service Business Services Authority

NOACs: novel oral anticoagulants

P2Y12i: P2Y12 inhibitors

PCSK9i: proprotein convertase subtilisin/kexin type 9 inhibitors

RAASi: renin-angiotensin-aldosterone systemic inhibitor

RCD: routinely collected data

RCTs: randomised controlled trials

SDTM: Study Data Tabulation Model

SGLT2i: sodium-glucose transporter 2 inhibitors

# Introduction

The purpose of this data analysis plan is to outline the objectives of the above-mentioned DPhil thesis,the outcomes used, and the statistical analyses planned. This DPhil degree started on October 2020 and will finish in September 2023.

# Background

## Rationale

Routinely collected data (RCD) offers many interesting opportunities for use in randomised controlled trials (RCTs) and other types of clinical research studies. Although RCD is increasingly being used for research, there are several challenges associated with its use. These include understanding the data features and structure, whether it is reliable and relevant1 to be used alongside - or in place of - traditional data sources in RCTs, and if its use is feasible.

During 2020, two new sources of RCD on medications in England have been made available to researchers: the General Practice Extract Service Data for Pandemic Planning and Research (GDPPR) dataset – which will be referred to as the GP dataset - and the Medicines Dispensed in the Community/NHS Business Services Authority (NHSBSA) dataset – henceforth referred to as the Dispensing dataset. Although the GP dataset is only available for COVID-19 research, it has been announced that a similar dataset is planned for release after September 2021 for other research purposes.

This thesis aims to assess whether these two new datasets concerning RCD on medications are a valid and feasible source of concomitant medication collection for RCTs. It will make use of data generated from three RCTs: RECOVERY, AMALFI, and ORION-4.

## Objectives

### Primary

The primary objectives will focus on comparing RCD with standard methods of data collection for answering prominent trial-specific questions of interest regarding concomitant medications in three large-scale RCTs, in order to determine which source is the best one in different scenarios; for this purpose, the following outcomes will be used:

* Immunosuppressive treatments at randomisation in the RECOVERY trial
* Initiation of anticoagulation after randomisation in the AMALFI trial
* Statin treatment intensity at randomisation in the ORION-4 trial

### Secondary

The secondary objectives will expand the comparative assessment of RCD versus standard methods (or between different RCD sources in RECOVERY) regarding broader patterns of concomitant medications of interest, namely:

* Representation of several drug groups of interest at randomisation using standard and RCD sources for participants in three RCTs (see section 4.3.6 for a complete list of drugs groups of interest)
* For RECOVERY, only RCD sources will be used as they are the only data collection method
* Initiation of antibacterial, antifungal, and antidiabetic drugs after randomisation for RECOVERY participants according to two RCD sources
* Initiation of rate- and rhythm-control agents after randomisation for AMALFI participants according to standard and RCD sources
* Total numbers of different drugs identified for ORION-4 participants from self-reported and RCD sources
* Representation of specific patterns of concomitant cholesterol-lowering medication use before and after randomisation for ORION-4 participants, according to standard and RCD sources (see section 4.3.3.2.3 for complete definitions)

### Subsidiary

The subsidiary objectives will focus on characterising the RCD sources themselves, with the aim of providing an in-depth insight into their structure, content, inter-relationships, and general potential,1 along with exploring additional assessments related to their use, namely:

#### RCD source features

* Description of the structure and content of two new sources of RCD on medications
* Investigation of the relationships between two sources of RCD for participants in the same trial

#### Clinical Data Interchange Standards Consortium (CDISC) alignment

* Feasibility of integrating RCD on medications into standard formats, namely the CDISC Study Data Tabulation Model (SDTM) Concomitant Medication (CM) domain in a large-scale RCT (ORION-4)

#### Continuous measures of medication adherence

* Application of methods to assess time-varying continuous measures of medication adherence and persistence from RCD

#### Methods to build SNOMED medication codelists

* Comparative performance of different methods to build medication codelists using the SNOMED terminology

#### Other exploratory assessments

* Evidence of previous statin use for statin-intolerant patients in ORION-4 (see section 7.5 for details)
* Relationship between statin use and LDLc levels at randomisation in ORION-4 according to different data sources (see section 7.5 for details)

Analyses planned for the subsidiary outcomes are detailed in section 7

## Study design

### Data collection

This study will include concomitant medication data from three RCTs currently being run at the Clinical Trial Service Unit (CTSU) within the Nuffield Department of Population Health at the University of Oxford, each of them using different sources.

#### RECOVERY (“Randomised Evaluation of COVID-19 Therapy”)

RECOVERY (NCT04381936) is a trial of treatments for hospitalized COVID-19 patients. RECOVERY has access to two sources of RCD on medications: the GP and the Dispensing dataset. The trial has regular monthly data drops from NHS Digital for both these datasets (starting in July 2020 for the GP, and December 2020 for the Dispensing dataset)

#### AMALFI (“Active Monitoring for Atrial Fibrillation”)

AMALFI (ISRCTN 15544176) is a trial of screening for subclinical atrial fibrillation in elderly ambulatory patients with a CHA2DS2VASc score of 3 or more in men and 4 or more in women.

The main source of medication data in AMALFI are extracts of primary care data performed by automatic searches at each of the general practices involved in the study (estimated 30-40). These are collected at 1, 2.5, and 5 years after randomisation (NB: for this thesis, only the 1 year data will be used).

Linkage data has been requested for the Dispensing dataset, and will be sought for the GP dataset if this becomes available.

#### ORION-4 (“A Randomized Trial Assessing the Effects of Inclisiran on Clinical Outcomes Among People with Cardiovascular Disease”)

ORION-4 (NCT03705234) is a phase-III trial of a cholesterol-lowering medication called inclisiran in patients with established cardiovascular disease who are 55 years or older.

The main source of concomitant medication data collection in ORION-4 are patient-reported medication lists, collected at each study visit on an eCRF by the local research teams. This includes a screening visit approximately 8-12 weeks before randomisation, a randomisation visit, a first follow-up visit at approximately 3 months, and subsequent follow-up visits at approximately 6-monthly periods thereafter (although follow-up visits can happen at any time after randomisation).

A linkage requested is currently (November 2021) being for the Dispensing dataset, and will be sought for the GP dataset if this becomes available.

### Overview of data sources

#### Routinely collected data sources

##### The GP dataset

The GP dataset is provided as a sub-product of the General Practice Extraction Service (GPES).2 GPES is a system operated by NHS Digital that regularly extracts data from GP practices in England for several purposes, namely reimbursement.

Rather than a complete extraction of all the information contained in primary care, GPES performs data extracts relating to specific clusters of interest, such as specific conditions, medications, and clinical care pathways. The GPES export comprises about 450 different clusters, which are built of approximately 50,000 individual SNOMED codes used in the extraction process.3 This dataset has no specified lookback period. Given the novelty of dataset this, its features and structure will be described as part of this project.

##### The Dispensing dataset

The Dispensing dataset is maintained and provided by NHS Digital using data generated by NHSBSA as a collection of electronic and paper prescriptions submitted for reimbursement each month (dating back to April 2018).4 This includes medicines dispensed or supplied by community pharmacies, appliance contractors, and dispensing doctors in England, as well as:

* medicines submitted by prescribing doctors and personally administered in England
* medicines prescribed in England but dispensed elsewhere in the UK
* medicines prescribed elsewhere in the UK but dispensed in England

The data includes prescriptions issued not only by general practitioners, but also community and hospital clinics, dentists, and community nursing services. Therefore, it is expected that this dataset will contain any medicine that was either prescribed anywhere in England, or dispensed in the community in England.

Similarly to the GP dataset, the features and structure of the Dispensing dataset will be described as part of this project.

#### Standard data sources

##### Local general practice data (AMALFI)

AMALFI is collecting data on medications and other features of interest from GP practices. Similarly to the GP dataset, the GP data extract in AMALFI is not a direct copy of all the data entered in the electronic health record, but based on a pre-specified search query developed by experienced primary care physicians and focused on medications of interest to the trial. This search is created by compiling clinical codes and terms that represent a specific feature of interest into a group of individual queries, each represented by a collection of clinical codes; the system then returns any records matching the codes of interest, grouped by query (e.g. aspirin, statins, insulins, etc). These data are direct extracts of the local record and can be considered prescription (rather than dispensing) data. The data fields collected include the medication name, date, units prescribed, number of issues in course, and instructions to patients.

NB: Although this is strictly speaking a RCD source, it will be considered as standard for comparison purposes since it was the established method used in this trial.

##### Patient-reported medication (ORION-4)

Concomitant medications for ORION-4 participants are recorded using electronic case report forms (eCRFs) at each study visit. All regular medications taken at the time of the visit are recorded using drop-down menus containing an in-house list of approximately 9,000 unique drug terms, or else by free text. The drug list is based on READ drug codes and additional bespoke drug codes assembled by researchers at CTSU, either for ORION-4 specifically or for previous trials. Each individual drug term has a unique in-house identifier (as well as a READ code if based on a READ concept). Free text entries are later assigned a pre-existing or new drug code. Doses are not recorded in ORION-4, except for statins.

### Definition of study outcomes

The main aim of this study is to assess how RCD compares with standard methods for concomitant medication collection in RCTs, and which sources of RCD are preferred. Therefore, the primary outcomes will replicate the main questions of interest regarding concomitant medications in RECOVERY, AMALFI and ORION-4; in doing so, these outcomes will depict the potential use of RCD in different disease settings and trial stages (baseline and follow-up).

Secondary outcomes will expand this assessment by exploring additional endpoints related to the primary outcomes and providing additional comparisons between data sources on broader categories of drugs in each trial. For secondary outcomes defined as treatment initiation, both proportions and time-to-event will be assessed to determine if any differences can be found between data sources when both a broad (proportion with a record in a specific time period) and strict (date of first record in that time period) definitions are used.

Additional analyses focused on subsidiary, data-centric assessments (see section 7) will expand the description of the RCD features, determine the feasibility of undertaking more granular assessments of adherence to concomitant medications using RCD, and explore the capacity for alignment of the RCD sources with established trial data standards.

#### Primary outcomes

* **Proportion of RECOVERY participants on background immunosuppressive treatments at randomisation,** according to each data source
* **Proportion of AMALFI participants starting an oral anticoagulant medication** (vitamin K antagonists or novel oral anticoagulants) within 1 year after randomisation (among those not taking them randomisation), according to each data source
* **Proportion of ORION-4 participants on each category of statin treatment intensity at randomisation** (as per the protocol definition, i.e. high-intensity: atorvastatin ≥40mg or rosuvastatin ≥20mg daily; low-moderate intensity: any regimen excluding high-intensity; and no statins), according to each data source

#### Secondary outcomes

##### RECOVERY

* **Total number and proportion of participants, cumulative number of records per participant, and time between records for each drug group of interest** (section 4.3.6.1) at randomisation and within 6 months after randomisation, according to each data source
* **Proportion initiating an antibacterial medication** within 6 months after randomisation, among those not taking them at randomisation, according to each data source
* **Time to first record of an antibacterial medication** within 6 months after randomisation, among those not taking them at randomisation, according to each data source
  + Note: for time-to-event assessments, a lag period of 7 days between prescribing and dispensing events will be defined and accounted for initially;5 an exploratory assessment will aim to establish a data-driven lag period for these RCD sources (see section 7.1, “Relationship between the GP and Dispensing datasets”)
* **Proportion initiating an antifungal medication** within 6 months after randomisation, among those not taking them at randomisation, according to each data source
* **Time to first record of an antifungal medication** within 6 months after randomisation, among those not taking them at randomisation, according to each data source
* **Proportion initiating an antidiabetic medication** within 6 months after randomisation, among those not taking them at randomisation, according to each data source
* **Time to first record of antidiabetic medication** within 6 months after randomisation, among those not taking them at randomisation, according to each data source

##### AMALFI

* **Total number and proportion of participants, cumulative number of records per participant, and time between records for each drug group of interest** (section 4.3.6.2) at randomisation and within 1 year after randomisation, according to each data source
* **Time to first record of an oral anticoagulant medication** (any of vitamin K antagonists or novel oral anticoagulants, and by these two separate groups) within 1 year after randomisation, among those not taking them at randomisation, according to each data source
* **Proportion initiating rate-control agents** (any of beta-blockers, calcium-channel blockers, and cardiac glycosides, and in each of these three groups separately) within 1 year after randomisation, among those not taking them at randomisation, according to each data source
* **Time to first record of a rate-control agent** (as defined above) within 1 year after randomisation, among those not taking them at randomisation, according to each data source
* **Proportion initiating rhythm-control agents** (any class I or III anti-arrhythmic drug, in each of these two groups separately, and amiodarone and flecainide only – see section 4.3.6.2), among those not taking them at randomisation, according to each data source
* **Time-to-first record of a rhythm-control agent** (as defined above), among those not taking them at randomisation, according to each data source

##### ORION-4

* **Total number and proportion of participants with a record of a drug group of interest** (section 4.3.6.3) at screening, randomisation, and each subsequent study visit (up until the 3rd follow-up visit), according to each data source
  + NB: cumulative number of records per participant and time between records for each drug group of interest will also be presented but for the Dispensing dataset only, as it is not possible to capture this information from the self-reported data
* **Number of different medications (active ingredients) identified per participant at each study visit** according to each data source
  + NB: for this assessment, each self-reported drug record will be assigned one or more WHODrug code (at the substance level) representing a single active ingredient, using an in-house procedure developed for ORION-4. In parallel, each entry in the Dispensing dataset (coded using SNOMED) will be assigned one or more active ingredient codes using the semantic features of the SNOMED terminology
* **Proportion identified as belonging to each statin intensity category at screening and each follow-up study visit** (as per protocol definition, see section 4.3.3.1, and section 4.3.9 for definitions of time points), according to each data source
* **Proportion initiating a statin of any kind after randomisation** (among those not taking a statin at randomisation), according to each data source
* **Time to first record of a statin for those initiating a statin of any kind,** among those not taking a statin at randomisation, according to each data source
  + NB: for self-reported data, the starting date will be assumed as the midpoint between the visit in which statins were first recorded and the last visit (see section 4.3.9.2); a sensitivity analysis will be performed using the date of the visit where statin use was first recorded
* **Proportion initiating ezetimibe after randomisation**, according to each data source
* **Time to first record of ezetimibe** not taking ezetimibe at randomisation, according to each data source
  + NB: the analyses method and sensitivity analysis will be the same as reported above for statins
* **Proportion initiating a fibrate** **after randomisation**, according to each data source
* **Time to first record of a fibrate** for participants not taking them at randomisation, according to each data source
  + NB: the analyses method and sensitivity analysis will be the same as reported above for statins
* **Proportion initiating a PCSK9 inhibitor after randomisation**, according to each data source
* **Time to first record of a PCSK9 inhibitor** for participants not taking them at randomisation, according to each data source
  + NB: the analyses method and sensitivity analysis will be the same as reported above for statins

### Main study hypothesis

For each of the primary and secondary outcomes, the null hypothesis is that the reference standard is a better representation of the true status of a participants’ concomitant medication. No hypothesis testing will be performed for subsidiary outcomes.

### Analysis populations and sample size

No sample size calculations have been performed for this study. A number of different populations will be included in this study, comprising participants of ongoing clinical trials who will continue to recruit during the study period; therefore the final number of individuals included in the study will only be determined at the end of August 2022 (allowing one year before the planned end of this DPhil course). The following populations will be considered:

#### RECOVERY

All patients randomised in England for whom linkage data from the GP and Dispensing datasets is provided to the study up until the end of July 2021

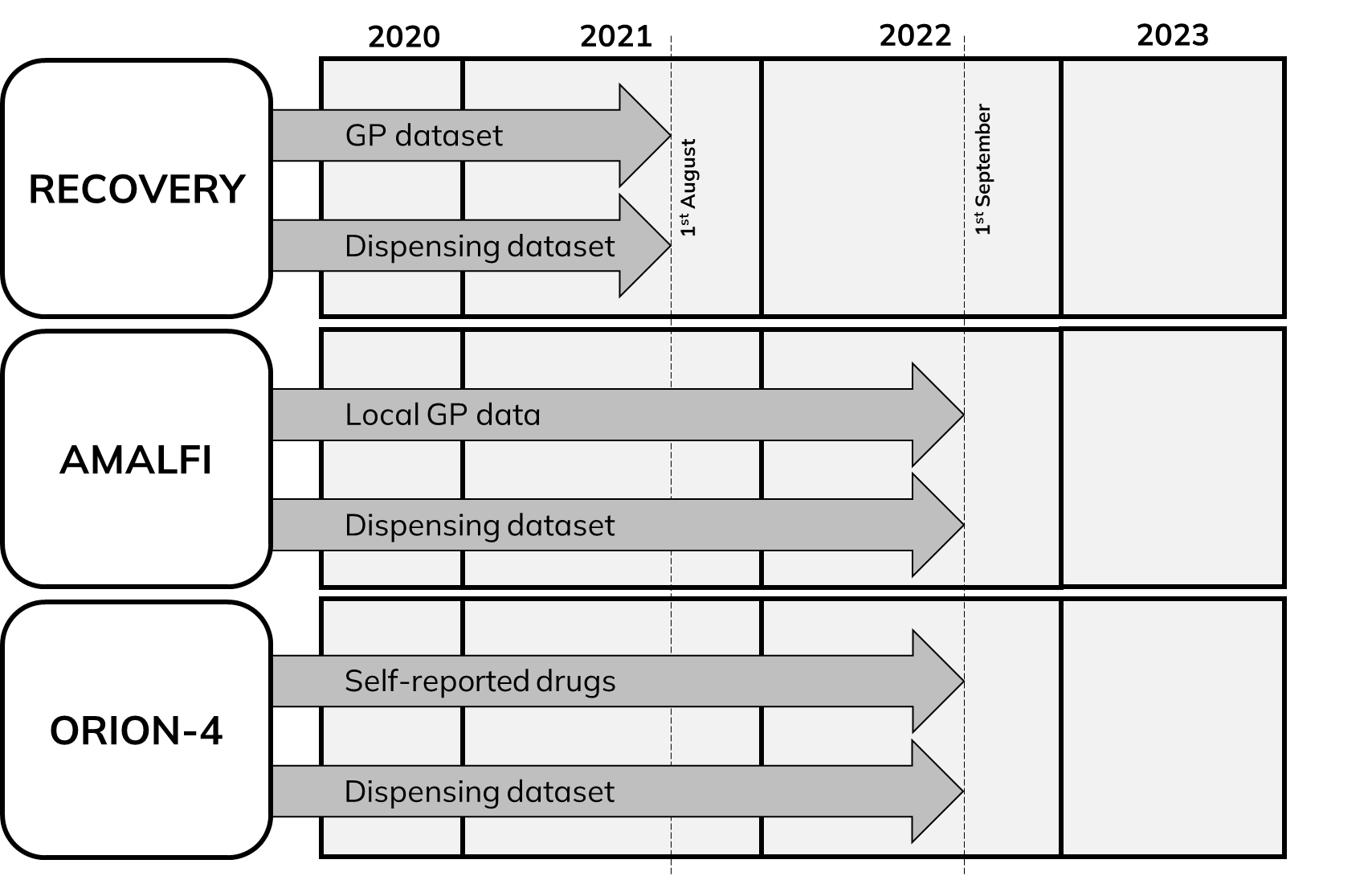
#### AMALFI

All randomised participants for whom both the 1-year local GP follow-up data and the linked Dispensing dataset could have been received by the end of August 2022 (i.e. a data extract from their practice has been received, and the linked extract from the cohort submitted to NHS Digital in which they were included has also been received, regardless of whether the extracts contain data relating to that patient)

#### ORION-4

All patients randomised in England for whom linkage data from the Dispensing dataset could have been received by the end of August 2022 (i.e. the linked extract from the cohort submitted to NHS Digital in which they were included has been received, regardless of whether the extract contain data relating to that patient)

Figure 1 - Data collection schedules

****

### Definition of drugs of interest

#### RECOVERY

RECOVERY has not pre-specified drugs of interest for its analyses. However, the following groups will be investigated in this study:

* Cardiovascular drugs:
  + Any cardiovascular drug
  + Any anticoagulant
  + Vitamin K antagonists
  + Novel oral anticoagulants (NOACs)
  + Low-molecular weight heparin (LMWHs)
  + Any antiplatelet
  + Aspirin
  + Dipyridamole
  + Clopidogrel
  + P2Y12 inhibitors (P2Y12i)
  + Any antiarrhythmic
  + Beta-blockers
  + Any calcium-channel blocker (CCBs)
  + Dihydropyridine CCBs
  + Non-dihydropyridine CCBs
  + Digoxin
  + Any antihypertensive
  + Any renin-angiotensin-aldosterone systemic inhibitor (RAASi)
  + Angiotensin-converting enzyme inhibitors (ACEi)
  + Angiotensin-receptor blockers (ARBs)
  + Mineralocorticoid receptor antagonists (MRAs)
  + Angiotensin-receptor/neprilysin inhibitors (ARNIs)
  + Renin antagonists
  + Other antihypertensives
  + Any diuretic
  + Loop diuretics
  + Thiazide diuretics
  + Alpha-adrenergic blockers
  + Other diuretics
  + Any anti-dyslipidemic
  + Statin
  + Ezetimibe
  + Proprotein convertase subtilisin/kexin type 9 inhibitors (PCSK9i)
  + Fibrates
  + Insulin
  + Metformin
  + Sodium-glucose transporter 2 inhibitors (SGLT2i)
  + Antidiabetic drugs (excluding insulin)
* Respiratory drugs:
  + Any respiratory drug
  + Anti-IgE antibodies
  + Cromoglycates
  + Histamine antagonists
  + Inhaled corticosteroids
  + Inhaled bronchodilators
  + Intravenous or intramuscular bronchodilators
  + Leukotriene antagonists
  + Nebulized drugs used in asthma or chronic obstructive pulmonary disease
  + Non-selective adrenoceptor agonists
  + Oral beta2-adrenoceptor agonists
  + Xanthines
* Antibiotic drugs:
  + Antibacterials
  + Antifungals
  + Antihelminthics
  + Antivirals
* Other drugs:
  + Antidepressants
  + Antipsychotics
  + Systemic immunosuppressive drugs (excluding corticosteroids)
  + Systemic corticosteroids

#### AMALFI

Initiation of anticoagulation (of any kind) is a pre-specified explanatory outcome in AMALFI and the main medication of interest to the study.

Other drugs of interest have been specified by the study team and are being extracted from primary care records, namely:

* + Aspirin
  + P2Y12 inhibitors
  + Beta-blockers
  + CCBs
  + Anti-arrhythmic drugs (any class I or III, and separately amiodarone and flecainide)
  + Digoxin or digitoxin
  + ACEi
  + ARBs
  + MRAs (spironolactone and eplerenone only)
  + ARNIs
  + Any renin-angiotensin system blocker (i.e. ACEi or ARB)
  + Any diuretic (excluding carbonic anhydrase inhibitors, mercurial diuretics, and osmotic diuretics)
  + SGLT2i (dapagliflozin, empagliflozin, canagliflozin, and ertugliflozin only)
  + Statins
  + Ezetimibe
  + PCSK9i
  + Fibrates
  + Insulin
  + Oral antidiabetic drugs
  + Injectable antidiabetic drugs
  + Nitrates
  + Vasodilator drugs

NB: the definitions of some these groups may vary from those used in the other studies

#### ORION-4

ORION-4 is collecting all concomitant medications reported by patients. These are grouped in the study database into 80 categories (including those derived for previous trials and not of specific interest to ORION-4). Based on this list, the following categories will be assessed:

* + Anti-dyslipidemic drugs
  + Any statin
  + Atorvastatin
  + Fluvastatin
  + Lovastatin
  + Simvastatin
  + Pitavastatin
  + Pravastatin
  + Rosuvastatin
  + Ezetimibe
  + PCSK9i
  + Fibrates
  + Resins
  + Antihypertensive drugs
  + Any antihypertensive
  + ACEi
  + ARBs
  + MRAs
  + ARNIs
  + Renin inhibitors
  + Beta-blockers
  + CCBs
  + Verapamil
  + Loop diuretics
  + Thiazide diuretics
  + Potassium-sparing and other diuretics
  + Alpha-blockers
  + Central antihypertensives
  + Vasodilator antihypertensives
* Antithrombotic drugs
  + Aspirin
  + Other antiplatelets
  + Vitamin-K antagonists
  + Heparin
  + Factor Xa inhibitors
  + Thrombin inhibitors
* Anti-anginal drugs
  + Nitrates
  + Other antianginals
* Antiarrhythmic drugs
  + Amiodarone
* Antidiabetic drugs
  + Insulin
  + Oral antidiabetic/hypoglycemic agents
  + Acarbose and similar agents
  + Biguanides
  + DDP4 inhibitors
  + Glinides
  + Glucagon-Like Peptide (GLP) 1 agonists
  + Sulphonylureas
  + SGLT2i
  + Thiazolideniones

### Derivation of drug groups and codelists

#### AMALFI

Drug groupings are based on the codes used to define each section in the primary care records extraction query, with outputs readily grouped according to each section (i.e. each retrieved record is located in a specific column within the data, which corresponds to a different drug group).

#### ORION-4

The ORION-4 data collection system contains 80 different drug groupings, which have been derived for ORION-4 or previous trials run in CTSU using the same drug coding system. These groupings are created by manual review and allocation of the drug codes available in the study database. No new groupings will be created for this project.

#### GP and Dispensing datasets

The two trials using standard drug collection methods (AMALFI and ORION-4) employ different recording systems, drug terminologies, and drug groupings. Similarly, the two RCD sources of interest here also use distinct data structures and drug terminologies (with no intrinsic drug groupings, although some groupings could be built based on the cluster system used to create the GP dataset).

As a consequence, although different trials might be interested in the same drug groups (e.g. antithrombotic therapies, such as antiplatelet drugs or anticoagulation), their definitions may vary - AMALFI has separate groupings for aspirin, P2Y12 inhibitors, vitamin K antagonists and novel oral anticoagulants (including factor Xa inhibitors and thrombin inhibitors), whereas ORION-4 has groupings for aspirin, other antiplatelets (including P2Y12 inhibitors and other less common drugs), warfarin-like anticoagulants, factor Xa inhibitors, and thrombin inhibitors. Similar cases apply for other therapeutic classes.

Additionally, initial exploration of the GP and Dispensing datasets as part of this study has highlighted that, although both use SNOMED as a drug coding terminology [with the Dispensing dataset also providing British National Formulary (BNF) codes], the hierarchical level used to record drugs is different – the GP dataset uses *product* terms such as “Aspirin 75mg”, while the Dispensing dataset uses *pack* terms such as “Aspirin 75mg pack of 28 capsules”. Since these are represented by different codes in the SNOMED terminology, the same codelist cannot be used for both datasets.

As a result, in order for meaningful comparisons to be made, separate codelists need to be built for each RCD source and for each drug group, so that they can provide a reasonable “match” of the intended groups used in the standard data collection methods.

Note: Although it is conceivably possible to build codelists that include codes representing a drug at all levels (using the semantic features of the SNOMED terminology), these are not currently available; as part of this project, an experimental method for this has been devised. This method will be compared with the main methods mentioned below in an exploratory assessment (see section 0)

In summary, based on the drug groups of interest for AMALFI and ORION-4, a matching codelist was built for use with the Dispensing dataset (the RCD source available for comparison in both trials) – this codelist represents the same conceptual drug grouping, rather than the same drug codes. Additional codelists were built for groups of interest to RECOVERY which were not covered by AMALFI and ORION-4 (such as respiratory, antibiotic, and immunosuppressive drugs). Then, separate matching codelists were built for the GP dataset to allow comparisons between the GP and Dispensing datasets for RECOVERY participants (and additional lists will be compiled to expand comparisons using the GP dataset for AMALFI and ORION-4 once this is made available for non-COVID purposes).

For the Dispensing dataset, the following procedure was used to compile codelists:

1. The codelist builder in OpenCodelists platform (www.opencodelists.org) was used to search the BNF classification for terms relevant to each codelist by one clinician
2. Matching terms were manually reviewed and selected
3. Where appropriate, previously derived codelists available in the OpenCodelists website were merged with the list and duplicates removed (based on the BNF code)
4. A second clinician reviewed and approved each codelist

For the GP dataset, the following procedure was used:

1. Manual review and allocation of the codes provided in the GDPPR cluster lookup list6 by one clinician
2. The codelist builder in OpenCodelists was used to search the BNF classification for terms relevant to each codelist by the same clinician
3. Matching terms were manually reviewed and selected
4. The resulting list was mapped to SNOMED codes using a standard NHS mapping5 integrated in the OpenCodelists platform, and merged with the list generated in step 1
5. Where appropriate, previously derived codelists available in the OpenCodelists website were merged with the list and duplicates removed (based on the SNOMED code)
6. A second clinician reviewed and approved each codelist

Besides drug groups of interest to each trial, separate codelists were built based on the BNF chapters (1-16) to allow high-level assessments of different therapeutic groups in both the GP and Dispensing datasets – for the GP dataset, the procedure followed the one described above (with no manual review by a second clinician). For the Dispensing dataset, no codelist was built since this dataset contains BNF codes whose structure allows interrogation based on the first digit of each code (which corresponds to the BNF chapter).

A complete list of all drug codelists built, and to which trial they refer, is presented in Table 12.

### Definition of reference standards

This thesis aims to assess the value of centralized RCD in comparison with established methods for medication data collection. Given the lack of an established reference standard for this purpose, for data relating to AMALFI and ORION-4 the respective “protocol” or standard data collection method will be used as a reference standard for comparison against the RCD source (see section 6.2).1

RECOVERY has not pre-established a preferred method of medication data collection, and the two data sources available for this purpose are both RCD. Therefore analyses of these data will not consider a reference standard and will instead investigate how the two sources relate to each other, and if one can be considered best (see section 6.2).

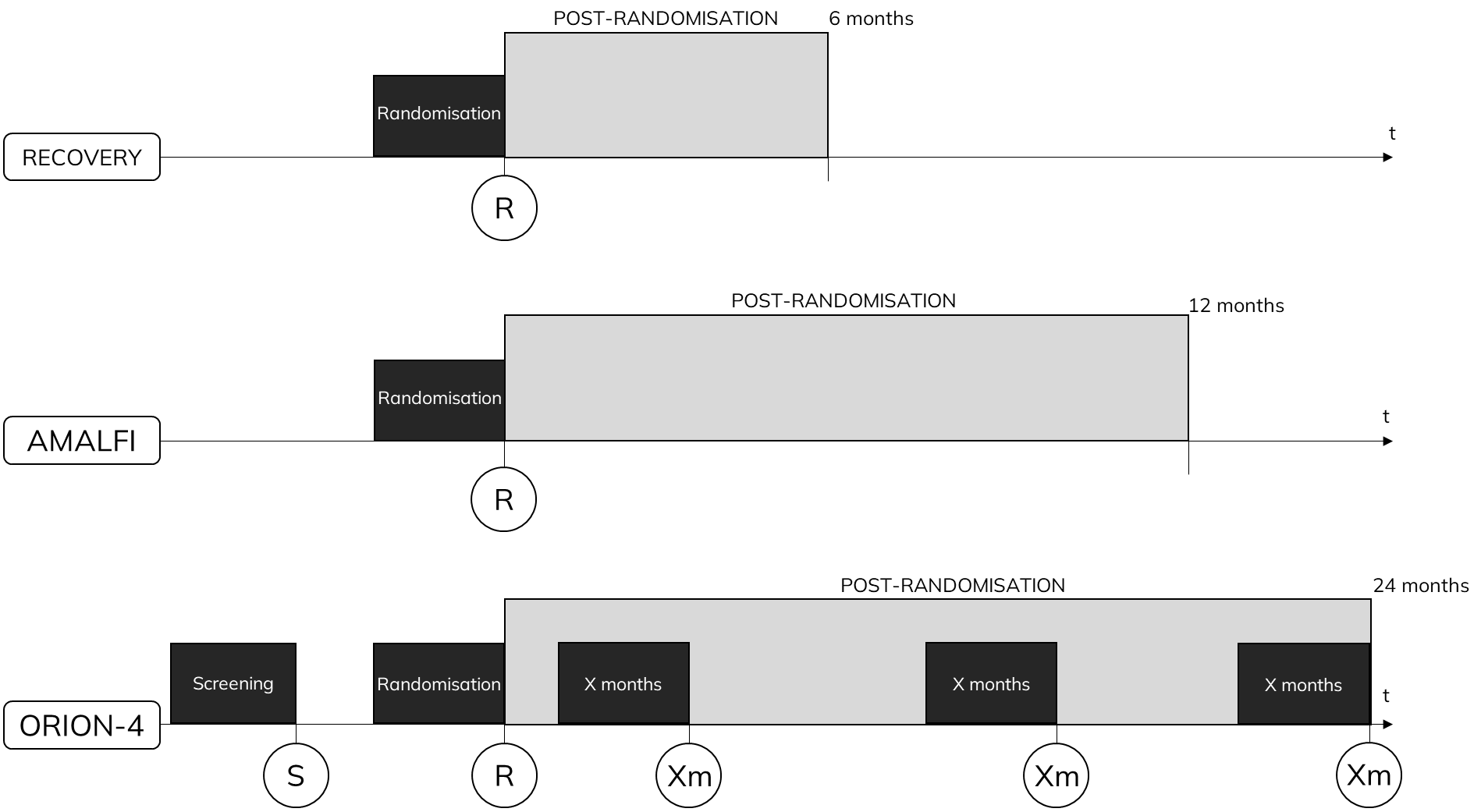
### Definition of time periods and time points

Concomitant medications can be recorded either as those a participant is currently taking at a single discrete time point (a study visit), or those that were taken during a time period of interest (exposure window).

Trials are typically interested in medications taken at or shortly before randomisation, and any changes that happen afterwards (during a specified follow-up period). Hence, this study will generally consider 1) a discrete pre-randomisation time point, and 2) a post-randomisation time period. Precise definitions will differ for each trial, based on specific protocol considerations, and additional time points will also be considered in ORION-4:

* RECOVERY and AMALFI do not have study visits, and are collecting information on concomitant medications from data extracts only (not at discrete time points) – therefore, for these two studies only a pre-randomisation time point will be considered, along with a post-randomisation time period (see **Error! Reference source not found.**Figure 2). The pre-randomisation lookback period will consider any record included within n days/months before the date of randomisation. The post-randomisation period will span 6 months for RECOVERY (a pre-specified time window for long-term analyses according to the study protocol, available at www.recoverytrial.net), and 12 months for AMALFI (the first time point for data collection from GP practices as per protocol, available at [www.amalfitrial.org](http://www.amalfitrial.org)); any record within these time periods will be considered.
* ORION-4 has several study visits at different time points (see **Error! Reference source not found.**), in which patient-reported medications taken at the time of the visit are recorded. These include screening, randomisation, and follow-up visits. Although the latter were intended to occur at approximately 3 months post-randomisation and 6-monthly thereon, ORION-4 patients can have multiple or no study visits in one pre-specified time point (per example, several visits may be made approximately at 3 months if a clinical concern arises in the first one; or a patient may miss their 3 months visit and only be seen later when the possibility arises, and at no particular time point – this is particularly true given the impact of the COVID-19 pandemic on face-to-face study visits). To avoid issues with assessing which study visits should be considered together, all visits happening within 24 months after randomisation will be assessed together, regardless of when they occurred. For each visit, a matching lookback period will be defined to allow interrogation of the RCD source (see section below)

Figure 2 - Schematic representation of data collection time points and analysis periods



Grey bars represent time periods; black bars represent time points

#### The lookback period

The lookback period represents a time window, lasting for n months (or days), in which medications taken by a patient are considered to be of relevance to their status at a particular contiguous and subsequent time point (such as randomisation), and are thus counted as a drug exposure. The exposure can be identified by a prescription or dispensing record which is considered to provide enough drug to last until the time point of interest, or by a record of a drug with an expected duration of action that spans that time point (although most drugs have relatively short effects of a maximum of 1-2 days, some drug classes might have lasting effects of weeks to months, such as systemic steroids, immunomodulatory agents, and PCSK9 inhibitors).

The lookback or exposure period does not have a standard definition, as this tends to be purpose and area specific. Two main approaches can be considered:7 1) an empirical or investigator-defined approach, in which an exposure period is defined based on clinical reasoning about usual patterns of prescription, or assumptions about the likely dosing – and therefore duration – of a prescription (see section 7.3), or 2) a data-driven approach that either uses a measure of “days of drug prescribed/dispensed”, if available, or aims to establish the likely duration of a prescription by taking into account the patterns of prescription and dispensing occurring in the real world and represented in the data – i.e. what is the average time period between each prescription or dispensing record for a particular drug or drug group.

Although an empirical approach is commonly used in trials and pharmacoepidemiological research (particularly when information on days of drug provided is not available), the feasibility of using a data-driven approach will also be assessed in this study as it is more objective and agnostic.7 The lookback period (for each study) will be determined as part of the secondary outcomes (see section 4.3.3.2); the overall average interval (in days) between prescription/dispensing records for each study will be used as the basis to define the lookback period in the first instance. If important differences arise between drug groups, distinct definitions may be used. The impact of different lookback periods will also be explored (see section 7.1).

For RECOVERY and AMALFI, time periods will be calculated using retrospective day counts starting from each patient’s date of randomisation (with records happening either before or after the period of interest being ignored).

For ORION-4, since visits can occur within a time window (and the comparison of interest is between data recorded at a visit and a concomitant “matching” RCD record), pre-visit periods will be calculated using retrospective day counts starting from each patient’s visit (rather than strict time points).

#### Definition of event dates

Initial exploration of the data sources available as part of this study has identified that each source provides dates associated with a medication record in different ways. The following dates will be used for analysis purposes:

1. For the local GP data in AMALFI, a single date is provided for each prescription record and will be the date used
2. For the patient-reported medication in ORION-4, only the date of the visit in which the drug is recorded is collected (i.e. there are no precise start or end dates). Thus, analyses concerning drug initiation or discontinuation will be performed using the halfway point between the last visit and the visit in which there is a medication change (as an approximation to the actual date)
3. For the GP dataset, two dates are provided for each record: a “date” field, and a “record\_date” field. A preliminary analysis will establish which of these is more likely to represent the actual date of the prescription (i.e. the earliest) versus the date in which that prescription was recorded, or if one field appears to be more reliable than other; the field deemed most accurate will be used thereon
4. For the Dispensing dataset, no precise dates are presented since all records are aggregated on a monthly basis, and only the month (“reporting period”) is provided. Thus, the first day of the corresponding month will be used as a surrogate for the actual dispensing date

# Descriptive analyses

## Participant throughput

For each trial, the flow of participants whose data has been included in the analyses presented will be described using a CONSORT diagram. This will describe the number of participants included in each study, those excluded due to withdrawal of consent, and those for which data is available from each data source (with no consideration of randomised allocations)

## Baseline characteristics of patient populations

Baseline characteristics of patients included in each trial will be presented separately for distinct populations (according to each trial’s data collection procedures):

1. RECOVERY

* All randomised participants
* Participants with linkage data available for both data sources
* Participants with linkage data available for the GP dataset
* Participants with linkage data available for the Dispensing dataset

1. AMALFI and ORION-4

* All randomised participants
* Participants with linkage data available for the Dispensing dataset

NB: If further data sources (namely the GP dataset) become available for any of these trials during the course of this study, additional linked populations will be defined and assessed

The characteristics assessed will depend for each trial and will be based on the trial’s baseline assessment (i.e. the randomisation form for RECOVERY, randomisation questionnaire for AMALFI, and screening form for ORION-4). These will include:

1. RECOVERY:

* Age
* Gender
* Days since symptom onset
* Days since hospitalization
* Level of respiratory support (none, supplemental oxygen or non-invasive ventilation, or invasive ventilation)
* Diabetes
* Heart disease
* Chronic lung disease
* Tuberculosis
* HIV
* Severe kidney impairment
* Ethnicity

1. AMALFI:

* Age
* Gender
* Heart failure
* Myocardial infarction
* Stroke or transient ischaemic attack
* Peripheral vascular disease
* Deep vein thrombosis or pulmonary embolism
* Diabetes
* Hypertension

1. ORION-4:

* Age
* Gender
* EQ5D quality of life scale (0-100)
* Smoking status
* Alcohol drinking habits
* Myocardial infarction
* Stroke or transient ischaemic attack
* Percutaneous coronary intervention
* Coronary artery bypass surgery
* Non-coronary arterial surgery or stent insertion
* Non-traumatic amputation
* Diabetes

# Comparative analyses

## Baseline characteristics

Baseline characteristics of the overall and linked populations (i.e. those for whom the RCD source is available) will be compared to inspect the presence of any biases related to how different populations may interact with the healthcare system and thus be represented in the RCD sources (e.g. underrepresentation of specific age groups, genders, ethnicities, geographies, etc).8,9 For AMALFI and ORION-4, comparisons will be made between the overall population and the linked population. For RECOVERY, comparisons will be performed for all randomised participants versus participants with linkage from each data source (i.e. two separate comparisons).

Continuous variables will be compared using independent samples T-test if normally distributed, or the Mann-Whitney U test if non-normally distributed; categorical variables will be compared using Pearson’s Chi-squared or Fisher’s exact tests where appropriate. Assessment of distribution normality will be performed by visual data inspection.

## Primary, secondary, and subsidiary outcomes

For comparisons of binary outcomes (with vs without record) between data sources, two-by-two contingency tables will be constructed for each trial and each drug group of interest. The measurement of agreement between data sources will be performed using Cohen’s Kappa;10 k >0.8 will be considered evidence of substantial agreement.11 Venn diagrams will be built depicting the number of participants captured by each source. Additionally, sensitivity, specificity, positive and negative predictive values will be computed based on a reference standard for each comparison (except for comparisons in RECOVERY; see section 4.3.8).1 In the absence of an established standard for this trial, the source which captures both more patients and more entries will be considered superior; if the data source deemed best varies for different drug groups this will be described.

For time-to-event outcomes (section 4.3.3.2), Kaplan-Meier curves will be constructed and the log-rank test will be used to test the null hypothesis of no difference between data sources.

Assessments of subsidiary outcomes will not include statistical testing, unless specified otherwise.

## Significance levels

Statistical significance thresholds will be set at alpha=0.05; two-tailed p-values and 95% confidence intervals will be presented where appropriate.

# Subsidiary analyses

Besides the primary and secondary outcomes, additional analyses will be performed to address the subsidiary (data-centric) objectives. These will include:

## RCD source features

* **Degree of completeness** (i.e. proportion of complete values) for each variable included in the GP (for RECOVERY) and Dispensing datasets (for RECOVERY, AMALFI, and ORION-4)
  + Overall and along time
  + For different age groups (RECOVERY: <50, 50-75, >75 years; AMALFI: <=75 and >75 years; ORION-4: <55, 55-75, and >75 years)
  + According to male and female gender
  + According to location where data was recorded
  + According to BNF chapter of the prescribed/dispensed medication
* **Total number of participants, cumulative number of records per participant, and average time between records for each BNF chapter** at randomisation according to each RCD source (i.e. GP and Dispensing datasets for RECOVERY; Dispensing dataset for AMALFI and ORION-4)
* **Average time between records for drugs according to drug formulation** (such as tablets, topicals, inhalers, ophthalmic/otic/nasal, injections, sublingual tablet and sprays, transdermal/suppositories, and liquids) in the GP (for RECOVERY) and Dispensing datasets (for RECOVERY, AMALFI and ORION-4)
* **Impact of variations in the duration and definition of the lookback period** (such as requiring one or more records for each group) on the proportion of participants identified in each concomitant medication group of interest (in RECOVERY, AMALFI, and ORION-4)
  + For ORION-4, receiver-operator-characteristic (ROC) curves will be built using different lookback durations and definitions, to assess which method most closely approximates self-reported medication at the randomization visit12
* **Variations in the number of prescriptions/dispensing records identified for each drug group of interest from RCD sources in different data extracts in the RECOVERY trial**
  + NB: for this assessment, an index extract will be identified and counts computed; then, the latest record date will be extracted and used to trim the two subsequent data extracts (so that they only contain records that could have been present in the index extract); three different extracts will be considered: the inception extract will be May 2021, and the subsequent ones June 2021 and July 2021
* **Correspondence between the GP and Dispensing datasets**:
  + Compare number of records per participant in the GP and Dispensing dataset for RECOVERY for particular drugs
    - This assessment aims to establish if there are records of Dispensing not showing in the GP data, and vice-versa
    - For this purpose, only records that could have been present in both datasets (i.e. are part of the GP extraction) can be used; for simplicity, single drug records such as aspirin and atorvastatin (rather than drug groups and combination drugs) will be used in this exploration
  + Assess the existence of a record identifier that allows linkage/matching of a prescription and dispensing record
    - This assessment aims to establish if there is a feasible way of matching a prescription with a dispensing event based on the data
    - If this is not the case, manual exploration of the data will be performed to establish if there are alternative matching options
  + Average time period between a prescription and the respective dispensing
    - This assessment will only be performed if a reliable matching method is devised (see above)
  + Number of dispensing events per prescription
    - This assessment will only be performed if a reliable matching method is devised (see above)

## CDISC alignment

* Development of datasets compliant with the Concomitant Medications (CM) Domain of the CDISC SDTM standards
  + This will include the description of which CM fields are required to capture the information contained in the RCD datasets, mappings between fields, description of transformations needed, and whether any useful additions could be made to the standards to accommodate these new sources
  + The alignment of the datasets produced with the standards will be assessed using the Pinnacle 21 application13

## Continuous measures of medication adherence

* Derive commonly-used continuous measures of medication adherence (Proportion of Days Covered – PDC, and Medication Possession Ratio - MPC) from RCD – as defined by Hess at al (2006)14
  + These measures require collection of data on “number of days prescribed/dispensed” or alternatively number of units dispensed and daily dose; if these are not provided, the following alternative methods will be explored:7,15
    - Researcher-defined methods:
      * Defining daily dose as the Defined Daily Dose (DDD)
      * Defining daily dose as one unit
      * Defining daily dose as the usual dose (based on the Summary of Product Characteristics)
    - Data-driven method
      * Defining number of days prescribed/dispensed as the average number of days between a prescription/dispensing record for that drug group in each dataset (the lookback period)
  + For each measure, both mean/median (and standard deviation or interquartile range) and proportion over 80% will be reported16–18

## Codelist development methods

In order to explore the performance of different potential methods for SNOMED codelist development (including a novel semantic-based method constructed as part of this project – see Appendix), comparative assessments will be made including the following methods:

1. Semantic method
2. Term-based method
3. BNF-SNOMED mapping (using OpenCodelists)
4. Rdiagnosislist (https://cran.r-project.org/web/packages/Rdiagnosislist/vignettes/SNOMEDcodelists.html)
5. Manual review and code selection from the GDPPR cluster reference table

These methods will be assessed separately for the GP and Dispensing datasets (except the manual review method, as it only applies to codes selected for inclusion in the GP dataset, and is therefore not expected to perform adequately in the Dispensing dataset). The assessments will cover: 1) number of unique codes included in each list; 2) number of patients identified; and 3) number of entries identified. No formal statistical testing will be performed for this analyses.

As a proof-of-principle, the following pre-specified drugs will be assessed: 1) aspirin; 2) atorvastatin; 3) salbutamol; 4) prednisolone; and 5) amoxicillin. These have been selected in order to provide a wide-range of medications used in different disease settings and with different formulations. Drug groups will not be used for simplicity, and since their definitions are highly-purpose specific (per example, different formulations of the same drug may need to be excluded depending on the aim). Moreover, selection of substances relevant to a particular drug group of interest can be easily performed using formularies and some clinical knowledge.

Appropriate statistical testing procedures will be devised and pre-specified in this plan before any comparisons are performed.

## Other exploratory assessments

* **Proportion of ORION-4 participants self-reported as not taking a statin at randomisation due to intolerance who have evidence suggestive of past statin use** (in the Dispensing dataset)
  + The purpose of this assessment is to explore the potential of RCD to verify previous use of statins in patients who self-report statin-intolerance (rather than define statin-intolerance)
  + For these patients, both number of dispensing records and number of different statins used will be captured
  + Although of clinical interest to ORION-4, this assessment has not been included in the secondary outcomes as it explores how different data sources can complement each other, rather than comparing them
* **Association of statin intensity** (three groups as per protocol: high-, low-moderate intensity, and none) **with LDL levels at randomisation for ORION-4 participants,** according to each data source
  + This assessment will be performed as biological validation of the RCD source in comparison with standard self-report
  + Two multiple linear regression models will be built (one for each source); independent variables will include age, gender, and statin intensity (categorized in three groups); the dependent variable will be LDL levels measured at randomisation (in mmol/L)
  + The model (and therefore source) with highest β and R2 will be considered superior

# Data monitoring

All data included in this study will be kept in secure and backed up servers maintained by the Nuffield Department of Population Health at the University of Oxford. This includes data from ongoing randomised trials. All analyses will be performed on copies of the data maintained by each trial team in the respective study database (mirror databases hosted in INGRES servers for AMALFI and ORION-4, and local copies of the study data for RECOVERY kept in dedicated and secure folders).

All scripts and outputs used will be saved in specific folders with restricted and password-protected access provided only to the author of this document and the supervision team.

# Software

R Studio version 1.4.1717 (or later) and R version 4.1.0 (or later) for Windows will be used for data retrieval, manipulation, analysis, and visualization.19,20

Data hosted in the AMALFI and ORION-4 study databases will be retrieved using SQL queries ran from within the R Studio environment using Open Database Connectivity (ODBC) connections, implemented through the *odbc* package.21

In addition to the base R language, the *tidyverse* package ecosystem22 will be employed for most tasks of data manipulation and visualization; this will include the *dplyr, dtplyr, lubridate,* and *stringr* packages for data manipulation, and the *ggplot2* package for data visualization. Other packages might be used as needed, and will be identified in the respective R script files

Additionally, the *adhereR* package23 will be used to calculate adherence measures (see section 7.3)

All code used to data analysis will be kept for inspection and reuse and hosted in GitHub at github.com/gpessoaamorim/DPhil\_RECOVERY

# References

1. Food and Drug Agency. FDA draft guidance for Industry: Real-World Data - Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. Published online September 2021.

2. NHS Digital. General Practice Extraction Service. Published September 13, 2021. Accessed September 21, 2021. https://digital.nhs.uk/services/general-practice-extraction-service

3. NHS Digital. GPES data for pandemic planning and research (COVID-19). Published September 20, 2021. Accessed September 21, 2021. https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research

4. NHSBSA. Dispensing data. Accessed October 26, 2019. https://www.nhsbsa.nhs.uk/prescription-data/dispensing-data

5. Mabotuwana T, Warren J, Harrison J, Kenealy T. What can primary care prescribing data tell us about individual adherence to long-term medication?-comparison to pharmacy dispensing data. *Pharmacoepidemiol Drug Saf*. 2009;18(10):956-964. doi:10.1002/pds.1803

6. NHS Digital. GPES Data for Pandemic Planning and Research - Guide for analysts and users of the data. Published September 20, 2021. Accessed September 21, 2021. https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data#code-clusters-and-content

7. Meaidi M, Støvring H, Rostgaard K, et al. Pharmacoepidemiological methods for computing the duration of pharmacological prescriptions using secondary data sources. *Eur J Clin Pharmacol*. Published online July 10, 2021. doi:10.1007/s00228-021-03188-9

8. Bohensky MA, Jolley D, Sundararajan V, et al. *Data Linkage: A Powerful Research Tool with Potential Problems*. Vol 10.; 2010:346. Accessed July 26, 2021. https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-10-346

9. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med Res Methodol*. 2014;14(1):36. doi:10.1186/1471-2288-14-36

10. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960;20(1):37-46. doi:10.1177/001316446002000104

11. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. 2005;37(5):360-363.

12. Curtis JR, Westfall AO, Allison J, Freeman A, Kovac SH, Saag KG. Agreement and validity of pharmacy data versus self-report for use of osteoporosis medications among chronic glucocorticoid users. *Pharmacoepidemiol Drug Saf*. 2006;15(10):710-718. doi:10.1002/pds.1226

13. Pinnacle21. SDTM Validation Rules. https://www.pinnacle21.com/validation-rules/sdtm

14. Hess LM, Raebel MA, Conner DA, Malone DC. Measurement of Adherence in Pharmacy Administrative Databases: A Proposal for Standard Definitions and Preferred Measures. *Ann Pharmacother*. 2006;40(7-8):1280-1288. doi:10.1345/aph.1H018

15. Qin X, Teng THK, Hung J, Briffa T, Sanfilippo FM. Long-term use of secondary prevention medications for heart failure in Western Australia: a protocol for a population-based cohort study. *BMJ Open*. 2016;6(11):e014397. doi:10.1136/bmjopen-2016-014397

16. Granger BB, Ekman I, Granger CB, et al. Adherence to medication according to sex and age in the CHARM programme. *European Journal of Heart Failure*. 2009;11(11):1092-1098. doi:10.1093/eurjhf/hfp142

17. Steiner JF, Prochazka AV. The assessment of refill compliance using pharmacy records: Methods, validity, and applications. *Journal of Clinical Epidemiology*. 1997;50(1):105-116. doi:10.1016/S0895-4356(96)00268-5

18. Karve S, Cleves MA, Helm M, Hudson TJ, West DS, Martin BC. Good and poor adherence: optimal cut-point for adherence measures using administrative claims data. *Curr Med Res Opin*. 2009;25(9):2303-2310. doi:10.1185/03007990903126833

19. R Core Team. A Language and Enviroment for Statistical Computing. In: ; 2017.

20. RStudio Team. RStudio: Integrated Development for R. In: ; 2020.

21. Hadley Wickham. Package ‘odbc.’ In: ; 2021. https://cran.r-project.org/web/packages/odbc/odbc.pdf

22. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *JOSS*. 2019;4(43):1686. doi:10.21105/joss.01686

23. Dima AL, Dediu D. Computation of adherence to medication and visualization of medication histories in R with AdhereR: Towards transparent and reproducible use of electronic healthcare data. *PLoS One*. 2017;12(4):e0174426. doi:10.1371/journal.pone.0174426

# Tables

Table 1 - Drug codelists

|  |  |  |  |
| --- | --- | --- | --- |
| **Codelist** | **RECOVERY** | **AMALFI** | **ORION-4** |
| **Cardiovascular drugs** | | | |
| Any cardiovascular drug | X |  |  |
| Any anticoagulant | X | X | X |
| Vitamin K antagonists | X |  | X |
| Novel oral anticoagulants (NOACs) | X |  |  |
| Factor Xa inhibitors |  |  | X |
| Thrombin inhibitors |  |  | X |
| Low-molecular weight heparin (LMWHs) | X |  | X |
| Any antiplatelet | X |  |  |
| Aspirin | X | X | X |
| Dipyridamole | X |  |  |
| Clopidogrel | X |  |  |
| P2Y12 inhibitors (P2Y12i) | X | X |  |
| Other antiplatelets (excluding aspirin) |  |  | X |
| Any antiarrhythmic (class I or III) | X | X |  |
| Amiodarone |  | X | X |
| Flecainide |  | X |  |
| Beta-blockers | X | X | X |
| Any calcium-channel blocker (CCBs) | X | X | X |
| Dihydropyridine CCBs | X |  |  |
| Non-dihydropyridine CCBs | X |  |  |
| Verapamil |  |  | X |
| Digoxin | X |  |  |
| Digoxin or digitoxin |  | X |  |
| Any antihypertensive | X |  | X |
| Any renin-angiotensin-aldosterone systemic inhibitor (RAASi) | X |  |  |
| Angiotensin-converting enzyme inhibitors (ACEi) | X | X | X |
| Angiotensin-receptor blockers (ARBs) | X | X | X |
| ACEi or ARB |  | X |  |
| Angiotensin-receptor/neprilysin inhibitors (ARNIs) | X | X | X |
| Mineralocorticoid receptor antagonists (MRAs) | X |  | X |
| Spironolactone or eplerenone |  | X |  |
| Renin antagonists | X |  | X |
| Other antihypertensives | X |  |  |
| Central antihypertensives |  |  | X |
| Vasodilator antihypertensives |  |  | X |
| Nitrates |  | X | X |
| Other antianginals |  |  | X |
| Any diuretic | X |  |  |
| Any diuretic (excluding carbonic anhydrase inhibitors, mercurial diuretics, and osmotic diuretics) |  | X |  |
| Loop diuretics | X |  | X |
| Thiazide diuretics | X |  | X |
| Alpha-adrenergic blockers | X |  | X |
| Other diuretics | X |  | X |
| Mineralocorticoid receptor antagonists (MRAs) | X |  |  |
| Any anti-dyslipidemic | X |  |  |
| Statins | X | X | X |
| Atorvastatin |  |  | X |
| Fluvastatin |  |  | X |
| Lovastatin\* |  |  | X |
| Simvastatin |  |  | X |
| Pitavastatin\* |  |  | X |
| Pravastatin |  |  | X |
| Rosuvastatin |  |  | X |
| Ezetimibe | X | X | X |
| Proprotein convertase subtilisin/kexin type 9 inhibitors (PCSK9i) | X | X | X |
| Fibrates | X | X | X |
| Resins |  |  | X |
| Insulin | X | X | X |
| Metformin | X |  |  |
| Sodium-glucose transporter 2 inhibitors (SGLT2i) | X |  |  |
| SGLTi (dapagliflozin, empagliflozin, canagliflozin, and ertugliflozin only) |  | X |  |
| Antidiabetic drugs (excluding insulin) | X |  |  |
| Oral antidiabetic drugs |  | X | X |
| Injectable antidiabetic drugs (other than insulin) |  | X |  |
| Acarbose and similar agents |  |  | X |
| Biguanides |  |  | X |
| Dipeptidyl-peptidase 4 (DPP4) inhibitors |  |  | X |
| Glinides |  |  | X |
| Glucagon-Like Peptide (GLP) 1 agonists |  |  | X |
| Sulphonylureas |  |  | X |
| Thiazolideniones |  |  | X |
| **Respiratory drugs** | | | |
| Any respiratory drug | X |  |  |
| Anti-IgE antibodies | X |  |  |
| Cromoglycates | X |  |  |
| Histamine antagonists | X |  |  |
| Inhaled corticosteroids | X |  |  |
| Inhaled bronchodilators | X |  |  |
| Intravenous or intramuscular bronchodilators | X |  |  |
| Leukotriene antagonists | X |  |  |
| Nebulized drugs used in asthma or chronic obstructive pulmonary disease | X |  |  |
| Non-selective adrenoceptor agonists | X |  |  |
| Oral beta2-adrenoceptor agonists | X |  |  |
| Xanthines | X |  |  |
| **Antibiotic drugs** | | | |
| Antibacterials | X |  |  |
| Antifungals | X |  |  |
| Antihelminthics | X |  |  |
| Antivirals | X |  |  |
| **Other drugs** | | | |
| Antidepressants | X |  |  |
| Antipsychotics | X |  |  |
| Systemic immunosuppressive drugs (excluding corticosteroids) | X |  |  |
| Systemic corticosteroids | X |  |  |

\* Lovastatin and pitavastatin do not have BNF codelists as they are not represented in the terminology; for the SNOMED codelists, an alternative method (not employing OpenCODELISTS) was used to interrogate the terminology

# Appendix

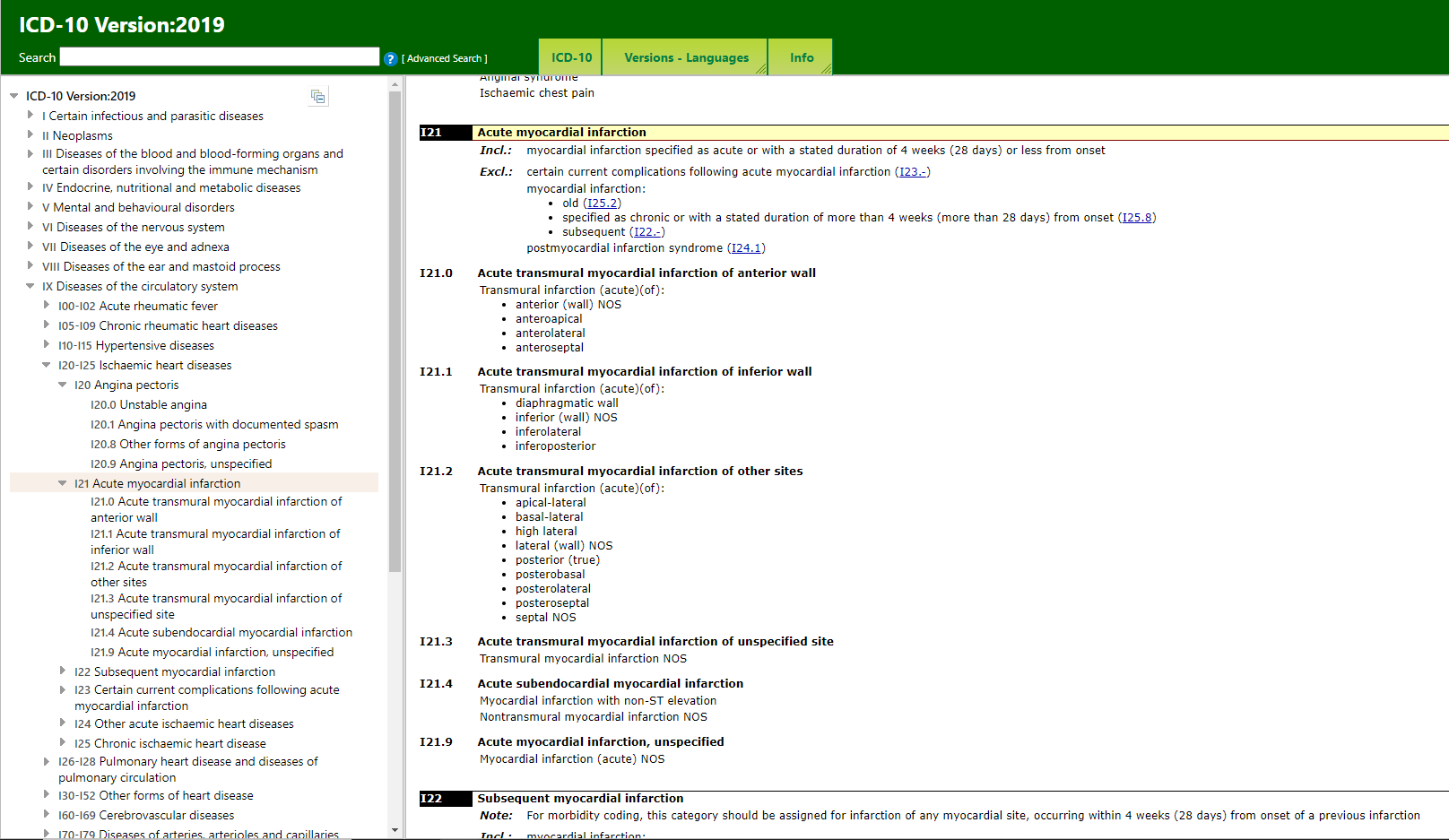
**Semantic-based SNOMED drug codelist development**

Clinical codelists (or codesets) depicting diagnoses, exposures, or drugs, for use in research, have typically been built using terminologies such as International Classification of Diseases (ICD) or BNF, which contain clinical diagnoses and medications, respectively. ICD-10 includes approximately 17,000 codes, while BNF includes approximately 50,000 codes; although not ideal, their sizes allow manual review and selection of codes if desired.

However, both these terminologies employ a mutually-exclusive hierarchical structure in which each concept can only belong to one parent level, and therefore has only one location possible in the hierarchical tree. Moreover, the code structure allows meaningful identification of different chapters and hierarchical levels within them (per example, all ICD-10 codes starting with “I” represent cardiovascular diseases, and all BNF codes starting with “02” represent cardiovascular drugs, with subsequent digits in each terminology representing subdivisions of these chapters – see Figure 3 - ICD-10 hierarchical structure

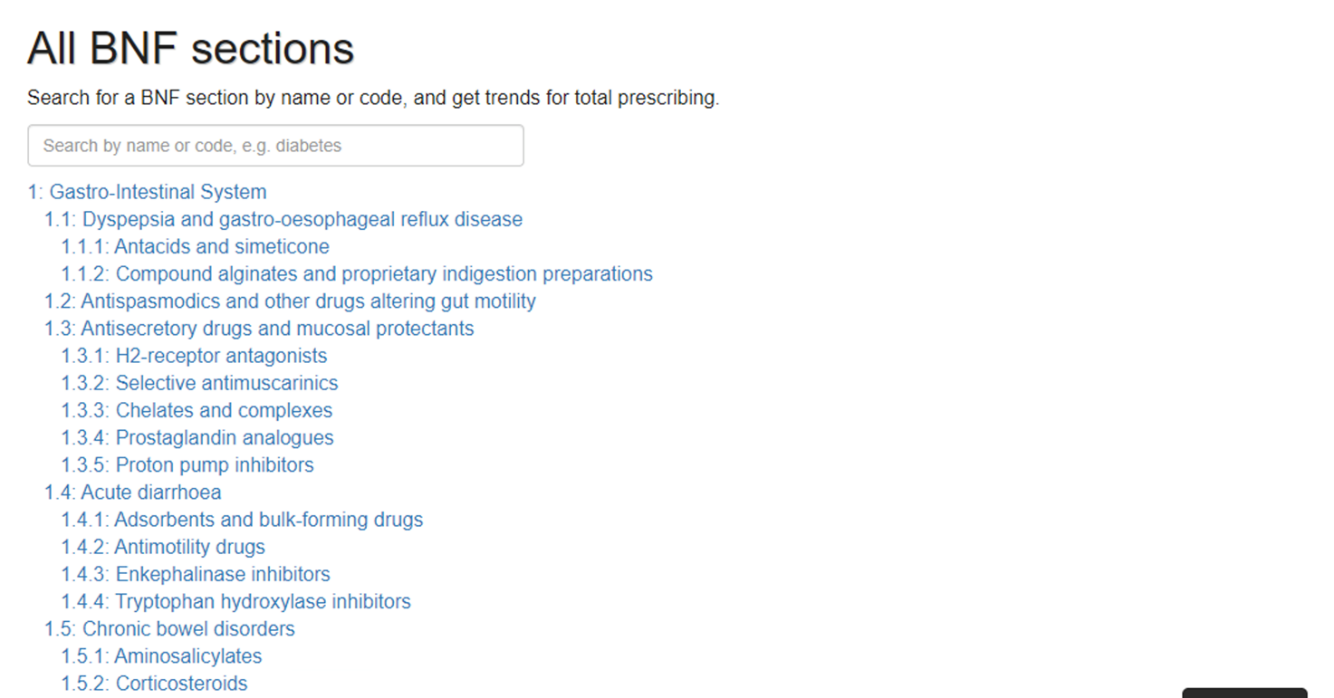
and Figure 4).

Figure 3 - ICD-10 hierarchical structure



Source: <https://icd.who.int/browse10/2019/en>

Figure 4 - Sample of the first BNF chapter

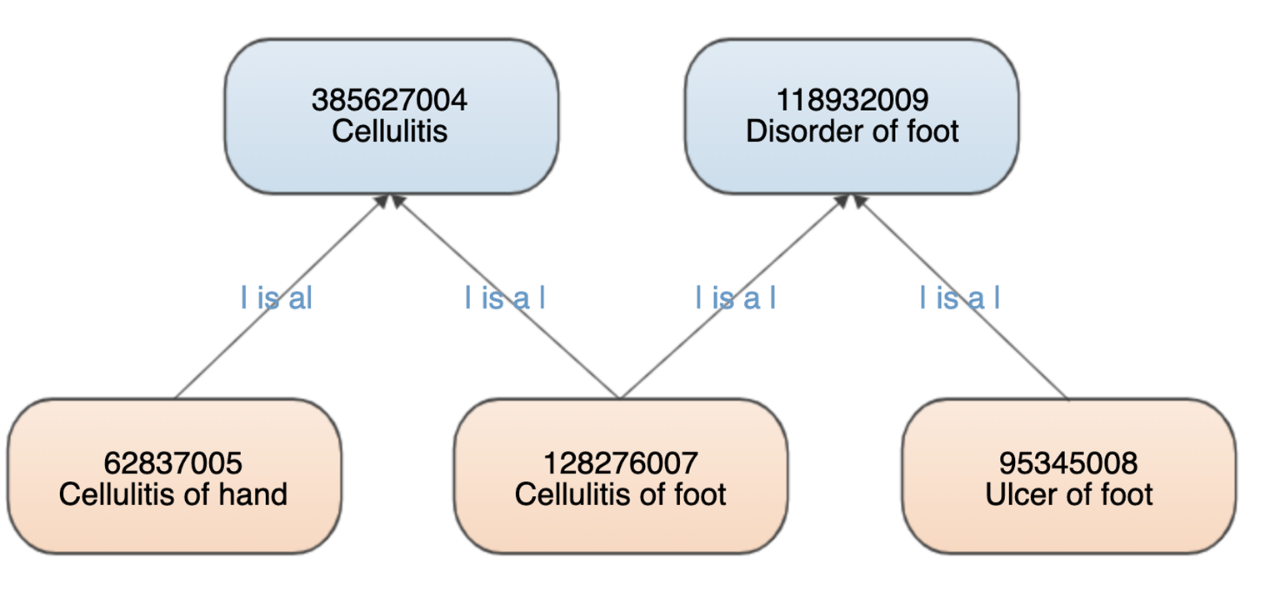


Source: https://openprescribing.net/bnf/

This structure makes code selection very straightforward, since all children of a level of interest may be included using string-based searches; per example, all ICD-10 codes contained between I20 and I25, regardless of the presence of additional digits, represent ischaemic heart diseases; similarly, all BNF codes starting with “0202” represent diuretics. For more general queries, higher level groups can be used, whereas for more detailed queries (such as including one particular diagnosis or a certain drug formulation) more terminal levels can be easily selected by following the hierarchical structure (together with string based searches of the terms in each terminology).

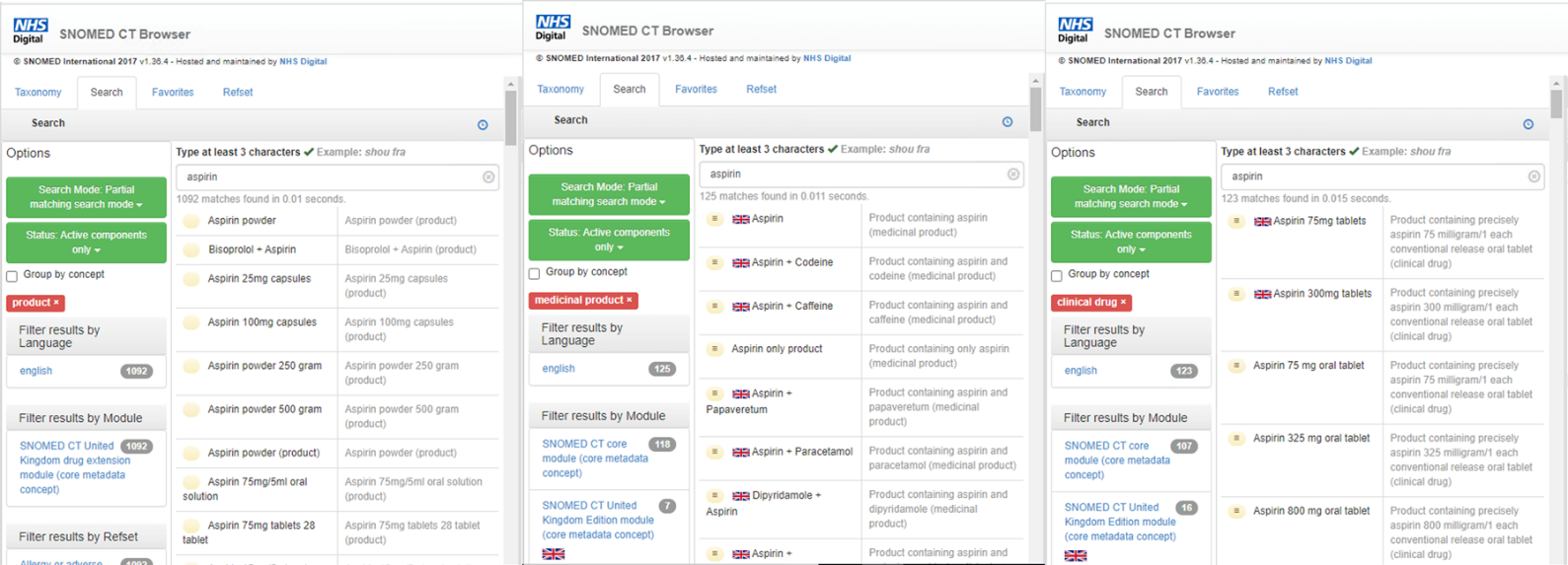
SNOMED-CT (the clinical data standard in the NHS) employs a different, multi-hierarchical structure, in which each concept may belong to multiple parent levels (see Figure 5) – prohibiting the use of hierarchical structures within the code itself, which is instead composed of a long numeric string with no particular meaning. It is also meant to capture all kinds of data relevant to healthcare, such as diagnoses, observations, laboratory results, anatomical structures, medications, etc. As a result, the total number of SNOMED codes is close to 2,000,000, making review of all codes an impractical task. Even if using the 19 higher-level hierarchical divisions (that separate drugs from diagnoses, per example), the multi-level hierarchical structure means that manual exploration of the hierarchy would take a substantial amount of time, and it would be quite difficult to track or audit – Figure 6 provides an example for aspirin drug codes.

Figure 5 - Example of the SNOMED multi-hierarchical structure



Adapted from: <https://confluence.ihtsdotools.org/display/DOCSTART/5.+SNOMED+CT+Logical+Model>

Figure 6 - Example of aspirin drug codes contained in different SNOMED hierarchies



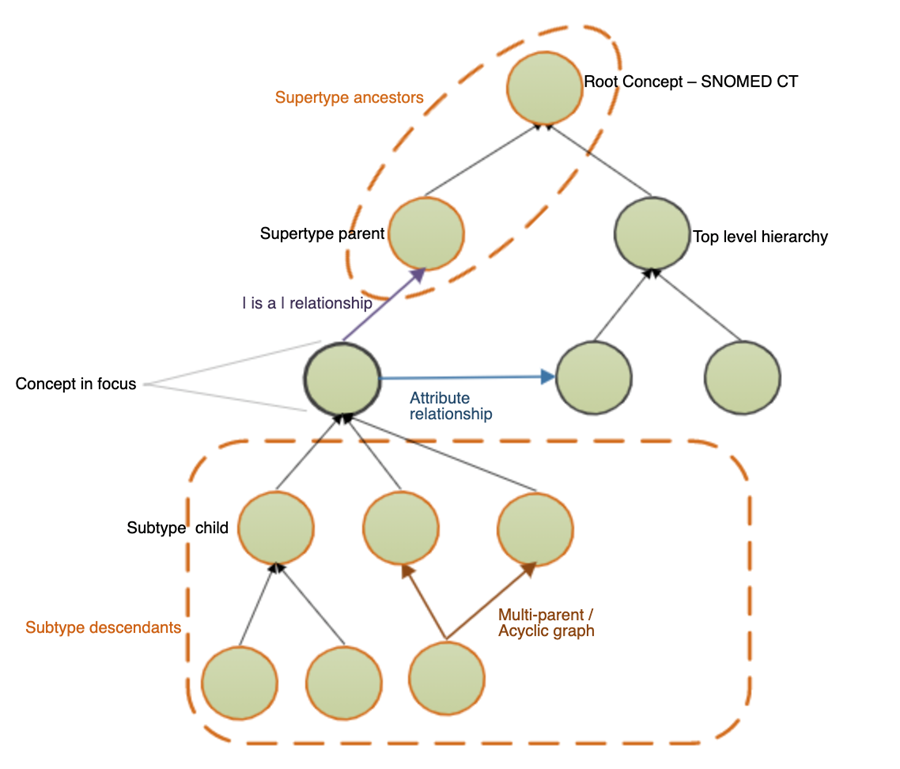
Aspirin drug codes are included in three separate hierarchies: product, medicinal product, and clinical drug

Moreover, whereas in BNF all commercial product names are aggregated as children of a higher term with the generic name (allowing string-based searches and then inclusion of all subconcepts), in SNOMED commercial names are not necessarily captured by string searches as they do not include the generic drug name, and their position in the hierarchy is not easily found by manual point and click approaches as for ICD-10 and BNF.

A potential solution to this problem lies in the semantic structure built in SNOMED-CT (see Figure 76 - Conceptual model of the SNOMED relational structure

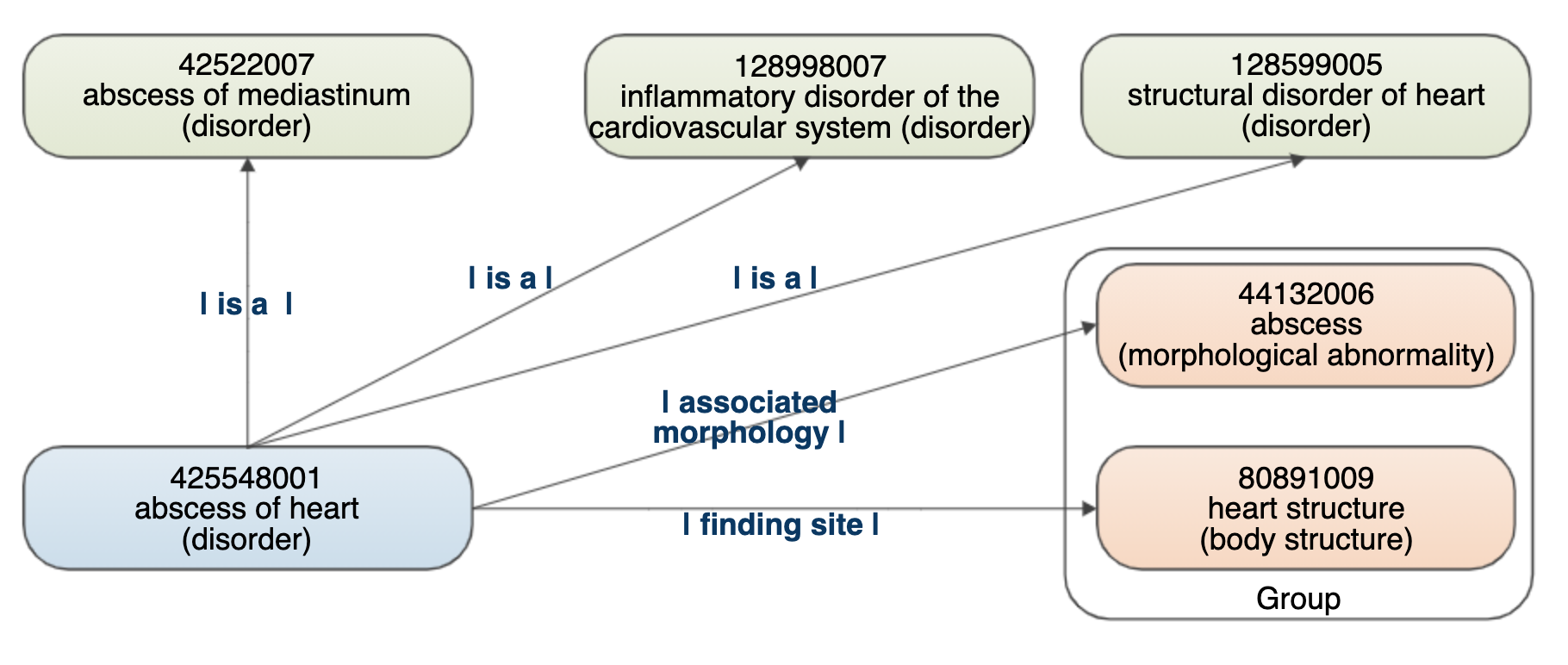
). Each concept in the terminology can be linked to other concepts via relationships. Whereas other terminologies such as ICD-10 and BNF also use (albeit indirectly) logical parent-children relationships (which specify where each concept lies within the hierarchy, and represented in the code itself), SNOMED provides an infinite number of possible relationships for each concept, such as “has finding site”, “has causative agent”, etc (see ). Therefore, apart from their position in the hierarchy (specified by “is a” parent-child relationship types), SNOMED concepts can have relationships specified with other concepts anywhere in the hierarchy.

Figure 7 - Conceptual model of the SNOMED relational structure



Adapted from: https://confluence.ihtsdotools.org/display/DOCSTART/5.+SNOMED+CT+Logical+Model

Figure 8 - Example of the SNOMED relational structure



Adapted from: https://confluence.ihtsdotools.org/display/DOCSTART/5.+SNOMED+CT+Logical+Model

This feature is particularly useful for assembling codelists, namely drug codelists. Besides “is a” relationships (SNOMED ceoncept 11668003), exploration of the terminology as part of this project has identified other relationships specific to drugs than can be used for this purpose, such as:

* + - 1. “has active ingredient” (127489000)
      2. “has AMP” [Actual Medicinal Product] (10362701000001108)
      3. “has basis of strength substance” (732943007)
      4. “has NHS dm+d (dictionary of medicines and devices) basis of strength substance” (10363001000001101)
      5. “has precise active ingredient” (762949000)
      6. “has specific active ingredient” (10362801000001104)
      7. “has VMP” [Virtual Medicinal Product] (10362601000001103)
      8. “is modification of” (738774007)
      9. “is pack of” (8652801000001103)

Selecting all codes for which a relationship of this type is specified and a destination concept contains the name of the drug of interest (e.g. “aspirin”) would therefore identify all potential variations of names of drugs that contain aspirin in their composition. This approach will be termed the “semantic” method for drug codelist assembly

A similar method is provided by the Rdiagnosislist package (for R), available at <https://cran.r-project.org/web/packages/Rdiagnosislist/vignettes/SNOMEDcodelists.html>. This package contains functions for loading the SNOMED terminology and assembling codelists, but only using “is a” relationships (i.e. selecting all children or all parents of a code of interest). The semantic method however theoretically more powerful as it is able to gather codes in distinct sub-hierarchies, and its added value can be assessed by comparing the number of codes retrieved by each approach, as well as the number of entries and participants identified in a single data source.