Logo

Description automatically generated

DATA ANALYSIS PLAN

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

DPhil thesis:

*Real-world drug data from electronic health records in the National Health Service:*

*validation and application in a randomized controlled trial of cardiovascular disease*

Version 1.0

Date: 16th August 2021

Author:

Guilherme Pessoa-Amorim

Supervisors:

Martin Landray

Louise Bowman

Will Stevens

Nuffield Department of Population Health

St Edmund Hall

University of Oxford

# Table of contents

[Table of contents 2](#_Toc83124888)

[1. Document history 3](#_Toc83124889)

[2. Abbreviations 3](#_Toc83124890)

[3. Introduction 4](#_Toc83124891)

[4. Background 4](#_Toc83124892)

[4.1. Rationale 4](#_Toc83124893)

[4.2. Objectives 4](#_Toc83124894)

[4.2.1. Primary 4](#_Toc83124895)

[4.2.2. Secondary 4](#_Toc83124896)

[4.2.3. Subsidiary 5](#_Toc83124897)

[4.3. Study design 6](#_Toc83124898)

[4.3.1. Data collection 6](#_Toc83124899)

[4.3.2. Overview of data sources 6](#_Toc83124900)

[4.3.3. Definition of study outcomes 8](#_Toc83124901)

[4.3.4. Main study hypothesis 11](#_Toc83124902)

[4.3.5. Analysis populations and sample size 11](#_Toc83124903)

[4.3.6. Definition of drugs of interest 11](#_Toc83124904)

[4.3.7. Derivation of drug groups and codelists 15](#_Toc83124905)

[4.3.8. Definition of criterion standards 17](#_Toc83124906)

[4.3.9. Definition of time periods and time points 17](#_Toc83124907)

[5. Descriptive analyses 21](#_Toc83124908)

[5.1. Participant throughput 21](#_Toc83124909)

[5.2. Baseline characteristics of patient populations 21](#_Toc83124910)

[6. Comparative analyses 23](#_Toc83124911)

[6.1. Baseline characteristics 23](#_Toc83124912)

[6.2. Primary and secondary outcomes 23](#_Toc83124913)

[6.3. Significance levels 23](#_Toc83124914)

[7. Additional analyses 24](#_Toc83124915)

[7.1. RCD source features 24](#_Toc83124916)

[7.2. CDISC alignment 25](#_Toc83124917)

[7.3. Continuous measures of medication adherence 25](#_Toc83124918)

[7.4. Other exploratory assessments 26](#_Toc83124919)

[8. Data monitoring 26](#_Toc83124920)

[9. Software 26](#_Toc83124921)

[10. References 28](#_Toc83124922)

# Document history

|  |  |  |
| --- | --- | --- |
| Version | Date | Changes |
| 0.1 | 21-09-2021 | Initial draft |

# 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 has started on October 2020 and will finish in September 2023.

# Background

## Rationale

Routinely collected data (RCD) offers many interesting opportunities for conducting clinical research, namely randomised controlled trials (RCTs). Although RCD is increasingly being used for research, there are several challenges associated with its use (especially when new datasets become available) – these include understanding the data features and structure, whether it is of sufficient quality 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 Prescribing dataset - and the Medicines Dispensed in the Community/NHS Business Services Authority (NHSBSA) dataset – henceforth referred to as the Dispensing dataset. Although the Prescribing 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 using RCD (compared to standard methods of data collection) to answer prominent trial-specific questions of interest regarding concomitant medications, as a use-case scenario in three large-scale RCTs, namely:

* 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 regarding broader patterns of concomitant medications of interest, namely:

* Representation of several drug groups of interest using both standard and RCD sources for participants in three RCTs
* 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
* Representation of specific patterns of previous and current statin use for ORION-4 participants according to standard and RCD sources

### Subsidiary

The subsidiary objectives will focus on answering questions about the RCD sources themselves, with the aim providing an in-depth insight into their structure, content, inter-relationships, and general potential, 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 participant 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)

#### Other exploratory assessments

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

## 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 Prescribing and the Dispensing dataset. The trial has regular monthly data drops from NHS Digital for both these datasets (starting in July 2020 for the Prescribing, and December 2020 for the Dispensing dataset)

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

AMALFI (ISRCTN15544176) 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 from the Dispensing dataset, and will be sought for the Prescribing dataset once this is made 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 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.

Linkage data has been requested for the Dispensing dataset, and will be sought for the Prescribing dataset once this is available.

### Overview of data sources

#### The Prescribing dataset

The Prescribing dataset is provided as a subproduct of the General Practice Extraction Service (GPES).1 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.2 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).3 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 Prescribing dataset, the features and structure of the Dispensing dataset will be described as part of this project.

#### Local general practice data (AMALFI)

AMALFI is collecting data on medications and other features of interest from GP practices. Similarly to the Prescribing 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.

#### 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-existent 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.

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 5) 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 2.3.6.1) 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
* **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 2.3.6.2) 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 2.3.6.2), 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 2.3.6.3) 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 each study visit** (as per protocol definition, see section 2.3.3.1), according to each data source
* **Proportion initiating a statin of any kind**, 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 2.3.9.2); 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** at 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 criterion 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 September 2022 (allowing one year before the planned end of this DPhil course):

#### RECOVERY

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

#### AMALFI

All patients for whom 1 year follow-up data is available from GP records until the end of December 2021.

#### ORION-4

All patients randomised in England for whom linkage data from the Dispensing dataset is provided to the study up until the end of December 2021.

### 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)
  + Anti-arrhythmic drugs (any class I or III – BNF section 2.3.2)
  + 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 – BNF section 2.3.2)
  + Amiodarone
  + 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

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.

#### Prescribing and Dispensing datasets

The two trials concerned in this project which are using standard drug collection methods (AMALFI and ORION-4) use 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 Prescribing 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 Prescribing 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 Prescribing 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 can 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.

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 (which is available for comparison in both trials) – i.e. a codelist that represents the same conceptual 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 Prescribing dataset to allow comparisons between the Prescribing and Dispensing datasets for RECOVERY participants (and additional lists will be compiled to expand comparisons using the Prescribing dataset for AMALFI and ORION-4 once this dataset 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 Prescribing dataset, the following procedure was used:

1. Manual review and allocation of the codes provided in the GDPPR cluster lookup list4 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 mapping 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 Prescribing and Dispensing datasets – for the Prescribing 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)

### Definition of criterion 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 gold-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 criterion standard for comparison against the RCD source (see section 4.2).

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 criterion standard and will instead investigate how the two sources relate to each other.

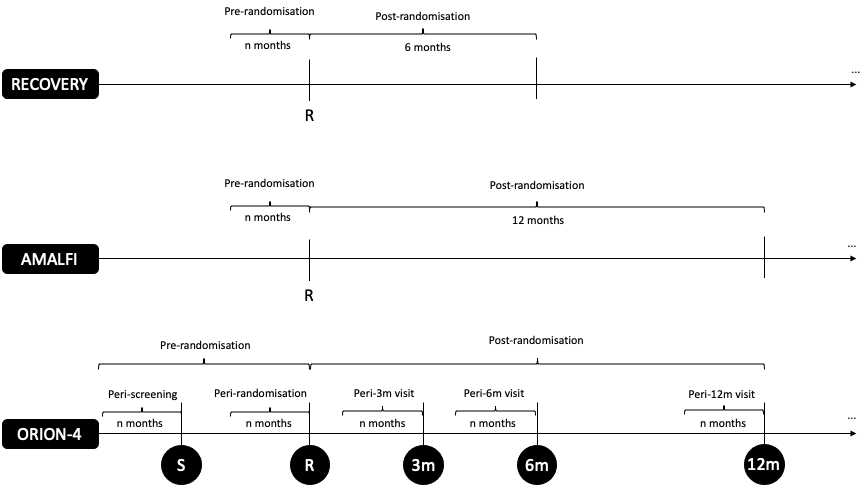
### Definition of time periods and time points

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 two distinct time periods: 1) pre-randomisation, and 2) post-randomisation. Precise definitions will differ for each trial, based on specific protocol considerations:

* 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, only time periods (rather than time points) can be considered for analysis (figure 1). 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). The pre-randomisation lookback period will span n months before the date of randomisation (and used as a representation of the patient status at the randomisation time point) – see below for details on the definition of the lookback period.
* ORION-4 has several study visits at different time points, in which patient-reported medications being taken at the time of the visit are recorded. To allow meaningful comparisons using RCD, a preceding “peri-visit” lookback period (lasting for n months) needs to be defined so that records occurring in that period can be used. For ORION-4, the post-randomisation period will be restricted to approximately 12 months (i.e. ending at the 3rd follow-up visit). This period will be separately defined for each patient since study visits do not have strict time windows.



**Figure 1 –** Schematic representation of data collection time points and analysis periods

#### The lookback period

The lookback period represents a time window, lasting for n months, 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:5 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 5.4), 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 data-driven approach will be the preferred method in this study as it is more objective and agnostic. The lookback period (for each study) will be determined as part of the secondary outcomes (see section 2.3.7.2); 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 5.1).

Besides the lookback period, additional time periods or time points may also be defined, such as the “run-in” period in ORION-4 (corresponding to the entire time period between the screening and the randomisation visit).

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), peri-visit periods will be calculated using retrospective day counts starting from each patient’s visit (rather than strict time points of 3, 6, or 12 months).

#### 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 Prescribing 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; the earliest date 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 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 Prescribing 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 Prescribing 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

For AMALFI and ORION-4, the baseline characteristics mentioned above will be compared between the overall population and the population for whom RCD is available (i.e. the “linked” population).

For RECOVERY, two separate comparisons will be made: 1) all randomised participants versus participants with linkage from any of the two data sources, and 2) participants with linked data from each data source.

## Primary and secondary outcomes

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.

For each trial and for each drug group of interest, two-by-two contingency tables will be constructed for comparisons of binary outcomes (with vs without record) according to different data sources. The measurement of agreement between data sources will be performed using Cohen’s Kappa.6 Additionally, sensitivity, specificity, positive and negative predictive values will be computed based on a criterion standard for each comparison (except for comparisons in RECOVERY; see section 2.3.8).

For time-to-event outcomes, 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.

## 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.

# Additional 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 Prescribing (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. Prescribing 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 Prescribing (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 visit6
* Variations in the number of prescriptions/dispensing records identified for each drug group of interested 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)
* Relationship between the Prescribing and Dispensing datasets:
  + Compare number of records per participant in the Prescribing and Dispensing dataset for RECOVERY for particular drugs
    - This assessment aims to establish if there are records of Dispensing not showing in the Prescribing data, and vice-versa
    - For this purpose, only records that could have been present in both datasets (i.e. are part of the Prescribing extraction) can be used; therefore, single drug records such as aspirin and atorvastatin will be used in this exploration
  + Assess the existence of a record identifier that allows linkage 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 (such as using date
  + Average time period between a prescription and the respective dispensing
    - This assessment will only be performed if a reliable matching method is devised
  + Number of dispensing events per prescription
    - This assessment will only be performed if a reliable matching method is devised

## 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 application7

## Other exploratory assessments

* **Proportion of ORION-4 participants self-reported as not taking a statin at randomisation due to intolerance who have evidence of trying at least 3 different statins in the past** (in the Dispensing dataset)
  + NB: 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

# 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.13,14

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.15

In addition to the base R language, the *tidyverse* package ecosystem16 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* package17 will be used to calculate adherence measures (see section 5.3)

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

# References

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

2. 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

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

4. 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

5. 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

6. 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

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

8. 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

9. Qin X, Teng T-HK, 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

10. 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

11. 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

12. 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

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

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

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

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

17. 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