Development, optimisation and implementation of artificial intelligence methods for real world data analyses in regulatory decision-making and health technology assessment along the product lifecycle

PI Prof. Dr. Britta Haenisch

Table of contents

Title: Work Package (WP) 1, Use Case 1 and 2

Pre-authorisation and Evaluation: Using Real World Data to Characterise a Representative Study Population of Patients Diagnosed with and Treated for *Breast Cancer* and *Amyotrophic Lateral Sclerosis (ALS)*

Date of last version of protocol	08-05-2023
Version	Final
EU PAS register number	
Active substance	NA
Medicinal product	Not specified
Product reference	Not specified
Procedure number	Not specified
Marketing authorization holder(s)	Not specified
Joint PASS No	No
The research question and overall objective	The research question: Can Real World Evidence from Real World Data (RWD) help to mitigate the limitations of traditional randomized clinical trials (RCT) and provide additional insights to traditional RCT in the pre-approval phase of medicinal products? The overall objective: The preparation of a good practice example for analyses of RWD for the pre-authorisation stage and the improvement of methods for external validity in observational data. Primarily, we will investigate the value of RWD from European national registers and statutory health insurance data (claims data) in generating high-quality, accessible, population-based information on breast cancer and ALS (diagnosis, treatment, outcome). Second, we will investigate the application of historical control arms of RCT neglected populations (for breast cancer) and synthetic data (for ALS) in the improvement of external validity as well as statistical power and precision.
Countries of study	Denmark, Finland, Germany, Portugal
Authors	Bräuner, Elvira Ehrenstein, Vera

Members of Work Packages 1-3 of the Real4Reg project and associates contributed to the protocol and adopt it:	Aborageh, Mohamed Adewuyi, Davis Arzideh, Roxana Becker, Cornelia Bräuner, Elvira Braithwaite, Billy Costa, Inês Ehrenstein, Vera Fernandes, Joana Froehlich, Holger Furtado, Cláudia Haenisch, Britta Hartikainen, Sirpa Heß, Steffen Horváth-Puhó, Erzsébet Kallio, Aleksi Korcinska Handest, Monika Roberta Nagy, Dávid Paakinaho, Anne Peltner, Jonas Pylkkanen, Liisa Roethlein, Christoph Schneider, Katharina Silva, Célia Tolppanen, Anna-Maijia Vancraeyenest, Aurélie Wicherski, Julia

Table of Contents for WP1, Use Case 1 and 2

1. List of Abbreviations

AE Adverse Events
AI Artificial Intelligence
ALS Amyotrophic Lateral Sclerosis
ATC Anatomical Therapeutic Chemical
BERT Bidirectional Encoder Representations from Transformers
CDM Common Data Model
CVD Cardiovascular disease
DFS Disease-free survival
ECOG Eastern Cooperative Oncology Group
EMG Electromyogram
ER Estrogen Receptor
EU European Union
HER2 Human Epidermal growth factor Receptor 2
HR Hormone Receptor
HRT Hormone Replacement Therapy
HTA Health Technology Assessment
ICD International Codes of Diseases
IRB Institutional Review Board
ML Machine Learning
MRI Magnetic Resonance Imaging
NOMESCO Nordic Medico-Statistical Committee
OC Oral Contraceptive
OMOP Observational Medical Outcomes Partnership
OS Overall survival
PCS Procedure Coding System
PFS Progression-free survival
PR Progesterone Receptor
RCT Randomised Clinical Trials
RWD/E Real World Data/Evidence
SES socio-economic status
STROBE Strengthening the Reporting of Observational studies in Epidemiology
TMN Tumour, Node, Metastasis classification
US United States
VNR Nordic Article number
WHO World Health Organisation
WP Work Package

2. Research question

2.1 Background - why the study is conducted

The use of real-world evidence (RWE) from real-world data (RWD) in drug development and regulatory decision-making is gaining traction. Although randomised controlled clinical trials (RCTs) have traditionally been the gold standard for generating clinical evidence, they have limitations making their results less generalisable to the entire target patient population. **RWE from RWD may help to mitigate these limitations and provide additional insights to traditional RCT in the pre- and post-approval phase of medicinal products.** Recent evaluations of marketing authorisation applications of new medicines show that RWE is present in all phases of drug development and considered as part of the authorisation application. The United States (US) Food and Drug Administration has issued draft guidance documents on data assessment, data standards, and the use of RWD, and Health Technology Assessment (HTA) Institutes are also developing frameworks on this topic. But the use of RWE from RWD in pre-authorisation and evaluation steps, including drug development, is still scarce and post-authorisation of RWD is often constrained. **To address this, the Real4Reg consortium combines expertise from regulatory bodies, health care data, academic expertise in RWE generation, statistical methods, supplementary expertise in Artificial Intelligence (AI)/Machine Learning (ML), and expertise in patient empowerment.**

Real4Reg is conducted to incorporate fit-for-purpose use case RWD selection, study design, and data analysis methodology, in doing this, Real4Reg seeks to potentially provide evidentiary value for the use of RWD in regulatory decision-making and HTA. Specifically, Real4Reg will attempt to provide AI/ML-based algorithms based on high-quality and representative data sources, including national healthcare registers and statutory health insurance data (claims data) from different European Union (EU) countries. This approach may enable heterogeneity and excellence at different levels of quality characteristics for the analysis of RWD, enabling methodology packages of high quality to meet the requirements of data analysis for regulatory purposes and HTA. In an attempt in harmonizing data, an Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) will be developed.

A total of six work packages (WPs) are included to address the overall aims in Real4Reg. In the present protocol, we will describe the specific objectives, data and methodology related to WP1 (Use Case 1 [generating a high-quality accessible population based common data model, describing disease, examination of heterogenicity between data sources from the four partners, development of computer programs for inclusion/exclusion criteria] and Use Case 2 [investigating the application of synthetic data and external historical control arms in the improvement of external validity and statistical power]). We will apply two

disease phenotypes, namely breast cancer (prevalent disease) and amyotrophic lateral sclerosis (ALS, rare disease) in WP1.

2.2 Overall objectives

The overall objective is the preparation of good practice example for analyses of RWD for the pre-authorisation stage and the improvement of methods for external validity in observational data. Primarily, we will investigate the value of RWD from European national registers and statutory health insurance data (claims data) in generating high-quality, accessible, population-based information on breast cancer and ALS (diagnosis, treatment, outcome). Second, we will investigate the application of historical control arms of RCT neglected populations (for breast cancer) and synthetic data (for ALS) in the improvement of external validity as well as statistical power and precision.

2.2.1 Specific aims for use case 1

Use case 1

For of all relevant exposure, covariate and outcome variables

1. Provide data access and carry out data pre-processing tasks including Quality Control and conversion to the Observational Medical Outcomes Partnership (OMOP) - common data model (CDM)
2. Provide a descriptive overview of covariates and meta data catalogue
3. Create a detailed data management plan including the assessment of data quality
4. Examine heterogeneity within the dataset, in the patient populations and in context in which the data are captured (differences in national healthcare registers and claims data) including:
    i. Coding practices in the four partners
    ii. Accessibility
    iii. Representativeness
    iv. Temporal variation in variables
    v. Bias
    vi. Completeness (missing data)

2.2.2 Specific aims for use case 2

Use case 2

We aim to extend our knowledge of AI/ML algorithms to the needs of data analysis for regulatory purposes by experimenting with AI/ML algorithms and synthetic data to examine and display which types of RWD can serve as high-quality external historical control arms/synthetic data.

Specifically:

2.2.2a Breast cancer

We will include descriptions of the treatment of breast cancer in otherwise RCT neglected populations to investigate validity. This will be done by creating historical population control arms including:

i. Pregnant women
ii. Women with co-morbidity that may influence participation including:
    a. psychiatric conditions
    b. cardiovascular disease (CVD)
iii. Older women (diagnosed at ages ≥65 years)
iv. Men with breast cancer
v. Women with poor performance status (ECOG ≥ 2)

2.2.2b ALS

We will create synthetic data.

2.3 The target population

The results from this present study may be generalizable to patients diagnosed with and subsequently treated for primary breast cancer or ALS in Europe. On a higher level the methodology packages of high-quality data produced in the overall Real4Reg study may potentially be applied for regulatory purposes and HTA.

2.4 Hypothesis

There is no *a priori* hypothesis for Use Case 1 and 2, WP1.

3. Study design

The overall first step in WP1 is to get an overview of all relevant covariates and their meta data. Afterwards, a detailed description of different variables will be central to illustrate the heterogeneity of disease data from four different data sources (Use Case 1). In addition, the context in which the data are captured will be considered after which we will experiment with AI/ML algorithms and synthetic data developed in WP3, to examine and display which types of RWD are able to serve as high-quality external historical control arms/synthetic data (Use Case 2).

3.1 Data collection

Data collection is secondary data.

Sources of data include:

1. Claims data registered in Germany from 2008 onwards
2. National healthcare registers from Denmark, Finland, and Portugal from 2000 onward.

3.2 Measures of occurrence

3.2.1 Breast cancer

For overall and subtype breast cancer the following measures of occurrence will be estimated:

1. Incidence rate (all and by stage at diagnosis, metastatic/non-metastatic) for primary breast cancer
2. Prevalence (all and by stage at diagnosis, metastatic/non-metastatic) for primary breast cancer

3.2.2 ALS

For ALS the following measures of occurrence will be estimated:

1. Incidence rate
2. Prevalence

3.3 Measure(s) of association

Not applicable. Adverse events/adverse reactions

3.4 Adverse events/adverse reactions

Not applicable.

4. Source and study population

4.1 Source population

Registrations of disease diagnoses according to WHO International Codes of Diseases, version 10 (ICD10) using health registers for the entire population in Denmark, Finland and Portugal and using claims records of public health insurance providers in Germany.

Details about data sources are presented in Table A.

4.2 Study population

4.2.1 Breast cancer

Persons diagnosed with incident primary breast cancer (ICD10/ICD-10-CM: C50) in the period from 2000 onward (Denmark, Finland, Portugal) and 2008 onward (Germany). The cohort defining criteria applied in each country are presented in Table B.

4.2.2 ALS

Persons diagnosed with incident ALS (ICD10 G12.2) in the period from 2000 onward (Denmark, Finland, Portugal) and 2008 onward (Germany). The cohort defining criteria applied in each country are presented in Table B. In Denmark, a specific code is available for ALS in the Danish adaptation of the ICD10, G12.2G. In Finland G12.2 is specific for ALS, other motoneuron diseases have other specific numbers. Portugal have their own specific number (ICD-10-CM G12.21) for ALS and Germany use a combination of disease code G12.2 and a dispensed prescription of riluzole (Anatomical Therapeutic Chemical [ATC] N07XX02), which is a treatment specific to ALS.

The duration of the follow-up will vary according to the outcome investigated (cf. section 5).

ALS can be hard to diagnose early because it can have symptoms similar to other diseases. Tests to rule out other conditions or help diagnose ALS (if available in the respective data sources used by each partner) might include:

1. Imaging and laboratory test:
    i. Magnetic Resonance Imaging (MRI)
    ii. Electromyogram (EMG)
2. Biopsies

4.3 Inclusion and exclusion criteria of the population

Table C and Table D describe the exclusion and inclusion criteria, respectively.

4.3.1 Breast cancer

The index date of the primary incident breast cancer is the date of the diagnosis in the cancer registry of Denmark, Finland and Portugal. In Germany, the index date is defined as the date of the first breast cancer diagnosis observable in the claims data.

The index date is considered the date of inclusion.

Inclusion criteria:

1. Diagnosis of a primary incident breast cancer
2. Age at time of diagnosis of 18 years or more

Exclusion criteria:

1. Prior history of other malignancy within the previous 5 years, except for carcinoma in situ of the cervix or basal cell carcinoma or squamous cell carcinoma of the skin that has been previously treated with curative intent
2. Not residing within the respective country with an active person ID number for at least 5 years prior to breast cancer diagnosis

Baseline details of breast cancer sub-type at diagnosis and the patient performance will be described (if available in the respective data sources used by each partner) including:

1. Hormone receptor status, HR+/-, estrogen receptor, ER, progesterone receptor, PR
2. HER-2 +/- status
3. Histology
4. Stage/TMN classification of malignant tumours
5. Invasion (yes/no)
    i. Grade for invasive breast cancer (G 1-3)
6. ECOG status

4.3.2 ALS

The incident ALS is the major inclusion criterion applied for Denmark, Finland and Portugal. In Germany, the index study inclusion date is defined as the date of enrolment in insurance coverage, without bridging of short gaps in enrolment.

Baseline is considered the date of inclusion.

Inclusion criteria

1. Diagnosis of incident ALS
2. Age at the time of diagnosis of 18 years or more

Exclusion criteria:

3. Not residing within the respective country with an active person ID number for at least 2 years prior to ALS diagnosis.

The wash-out period for ALS medication is a period of a minimum of 2 years (maximum of 5 years), prior to the start of the observation period and persons with riluzole medication (ATC N07XX02) are excluded. This wash-out period will vary according to available data in each country and heterogenicity will be reported.

Baseline details of ALS will be described (if available in the respective data sources used by each partner) including:

1. Histology
2. Prodromal stage is 5 years before the ALS diagnosis
3. Initial treatment

5. Treatment

5.1 Definition of treatment

All treatments will be ascertained using WHO ICD10/ICD-10-PCS procedure codes and ATC classification codes in the available registers and claims data and are described in Table G.

5.1.1 Breast cancer

Breast cancer treatments include:

1. Surgical treatment
    i. Breast cancer resection vs. mastectomy
        - if mastectomy, immediate breast reconstruction (yes, no)
    ii. Sentinel node biopsy (yes, no)
    iii. Lymph node dissection (yes, no)
2. Radiotherapy
3. Chemotherapy, including type and length of chemotherapy if available in the specific registry:
    i. Neoadjuvant therapy
    ii. Adjuvant therapy
        a. Anthracycline-based regimens
        b. Taxane-based regimens
        c. Other regimens (including platinum-based regimens, please specify)
    iii.
    iv. Treatment of metastatic disease
4. Hormonal treatment
5. HER-2 -directed therapy
6. Other treatment

5.1.2. ALS

ALS treatments include (if available in the respective data sources used by each partner prescription drug, treatment with riluzole (ATC code: N07XX02).

5.2 Validity of the treatment measurement

Treatment measurements of breast cancer will be based on NOMESCO, ICD10~~,~~/ICD-10-CM/ICD-10-PCS codes [procedure], VNR [medicine brand names, active ingredient, strength of the medicine, package size], ATC code [medicines] and visit date/visit type [secondary care, in/out/ER visits]. Treatment measurements of ALS will be based on ATC-codes (main treatment), VNR [medicine brand names, active ingredient, strength of the medicine, package size], and NOMESCO, ICD10,/ICD-10-CM/ICD-10-PCS codes for MRI and EMG procedures.

5.3 Time windows of treatment

Treatment is assessed after inclusion into the cohort with a baseline primary incident breast cancer diagnosis/with a baseline incident ALS diagnosis as a time-varying variable. Different time windows will be considered for treatment definitions.

5.4 Intensity of treatment

Dates of breast cancer/ALS treatment will be considered and duration of treatment will be derived from this. We will not consider treatment dose.

5.5 Biological mechanism of treatment

Not applicable

5.6 Comparators

5.6.1 Breast cancer

We will include descriptions of the treatment of breast cancer in otherwise RCT neglected populations. This will be done by creating historical population control arms including:

1. Pregnant women
2. Women with serious co-morbidity that may influence study including:
    i. psychiatric conditions
    ii. cardiovascular disease
3. Elderly women (diagnosed at ages ≥65 years)
4. Men with breast cancer
5. Women with poor performance status (ECOG ≥ 2)

In WP1, the external validity for breast cancer is investigated in Use Case 2.

5.6.2 ALS

We will experiment by with AI/ ML algorithms and synthetic data, developed in WP3.

In WP1, the improvement of statistical power and precision of the investigations related to ALS, is investigated in Use Case 2.

5.7 Baseline descriptive data on predefined covariates

Using European national registers and statutory health insurance data (claims data) we will characterize the b*aseline population* of identified patients.

Table E describes details of all relevant available predefined covariates that will be applied to address descriptive analysis in Use Case 1 and 2, stratified by country.

5.7.1 Breast cancer

We will include the following covariates, measured at or before the index date with an appropriate lookback period, where available:

1. Demographics/characteristics
    i. Age
    ii. Sex
    iii. Socio-economic status, SES (highest level of education, income if available, or a proxy SES variable such as zip code)
2. History of oral contraceptive (OC) use
3. History of hormone replacement therapy use (HRT, estrogen only, progesterone only, combination)
4. Co-morbidities (including psychiatric and CVD diseases)

In addition, for breast cancer we include information on parity, oophorectomy, mastectomy and breast reconstruction (preservation). These three variables are not included on the predefined Table E covariate list.

5.7.2 ALS

We will include the following baseline covariates, where available:

5. Demographics/characteristics
    iv. Age
    v. Sex
    vi. Socio-economic status, SES (highest level of education, income if available, or a proxy SES variable such as zip code)
6. Co-morbidities (including psychiatric, respiratory, and cardiovascular diseases)

Section 5 described the selection of covariates and potential confounders.

6. Outcomes and follow-up

Outcomes are described in Table G.

Outcomes will be ascertained using signals for disease progression. More specifically, outcomes will be ascertained using secondary data registered in national healthcare registers and claims data. Definitions are based on using ICD10 diagnosis/procedure codes registered in hospital registers and causes of death in causes of death registers.

Use Case 1 is largely descriptive and does not consider follow-up.

In Use Case2 we utilize complete follow-up from the date of diagnosis (i.e. the index date of BC* or ALS separately).

The follow-up, depending on the outcome will end on the earliest of

1. the event of interest
2. death (unless it is itself the event of interest)
3. emigration/disenrollment from the database
4. administrative censoring on 31 December 2021

* The index date may be different for e.g., the population of nonmetastatic and metastatic BC.

6.1 Breast cancer

The outcomes selected are those used in major RCTs in breast cancer. These also include the key outcomes on which regulatory approvals of new drug products in breast cancer are based. In addition, outcomes used in epidemiological studies are listed.

6.1.1 Primary outcomes

The following include the key outcomes for primary breast cancer in clinical studies for regulatory approval:

1. Overall survival (overall, and at 5-year and at 10- years)
2. Breast cancer-specific survival (overall, and at 5-years and 10-years)
3. Disease-free survival (DFS) for non-metastatic disease
4. Progression-free survival (PFS) for metastatic disease
5. Change in frequency of hospital visits and primary care visits before and after BC treatment

The following include the key outcomes in epidemiological studies:

6. Incidence rate (all, metastatic, non-metastatic)
7. Point (end of follow-up) and 10-year period prevalence (all, metastatic, non-metastatic)

8. Mortality (all-cause mortality and breast cancer -specific mortality, if available)

All outcomes will be standardised to the European Standard Population to enable comparisons between the four data nodes. https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-ra-13-028

6.1.2 Secondary outcomes

1. Safety outcomes (adverse events [AEs], Grade 3-4 AEs, serious AEs, deaths during treatment, long-term adverse events)
2. Rate of new morbidities
    i. CVD
    ii. Respiratory disease
3. Rate of secondary cancers (not breast cancer, C50*)

6.2 ALS

6.2.1 Primary outcomes

1. Incidence rate
2. Point (end of follow-up) and 10-year period prevalence
3. ALS- specific survival
4. Overall survival (OS)
5. Change in frequency of hospital visits and primary care visits before (prodromal stage) and after diagnosis

6.2.2 Secondary outcomes

1. Mortality rate
    i. All-cause
    ii. Disease specific (particularly respiratory failure) if available.
        2. Time interval from disease diagnosis to death
        3. Rate of new comorbidities
    iii. Cardiovascular disease
    iv. Respiratory disease

6.3 Validity of assessment of outcome measurement from healthcare databases

Certain outcomes, such as survival and mortality and the causes of death can be derived directly from registers or healthcare databases. New primary cancers during follow-up can be identified from cancer registers. In addition, information of disease status (non-metastatic/metastatic) may be obtained by the use

of certain healthcare codes indicating “metastatic disease” (e.g. for Denmark) or “treatment due to metastatic disease” (e.g. for Finland). Unfortunately, these codes have been used only during the last few years, and the data are incomplete.

However, not all-important outcomes are derived directly from registers, but they can be ascertained by using secondary data registered in national healthcare registers and claims data.

Some outcomes may be identified by using signals from utilization of healthcare resources. For example, as a proxy for disease progression, special healthcare visits in out-patient care with breast cancer codes, referrals from out-patient care to specialized breast cancer services with breast cancer codes, referrals to palliative care units with breast cancer codes, results of imaging examinations [e.g., mammography, CT, MRI] and laboratory findings suggestive of cancer recurrence, can be used. Furthermore, the start of new treatment (identified e.g., by codes for surgical procedures, radiotherapy or chemotherapy after several years without these kinds of visits, or start of new medication for breast cancer from registers) can be used as a proxy for disease recurrence/progression. However, this kind of estimation comes with a lot of uncertainty.

In general, we will attempt to increase external validity by including otherwise neglected populations (cf. section 5.6, comparators).

Validity relating specifically to outcome measurements is expected to be varied. The registration of disease, medical procedures, death and causes of death is mandatory in Denmark, Finland and Portugal and these ICD codes follow WHO classification and are validated and documented centrally. Furthermore, healthcare is free in for example Denmark and Finland, implying little ascertainment bias. The Portuguese National Cancer registry has full coverage of public and private sectors; however, in relation to other episodes of healthcare, the Portuguese National Health Service only collects data from public entities. Therefore, there is no data about healthcare delivery in the private sector, when citizens opt to go that way. In Germany, mandatory health insurance and social programs are available to pay for people without income that are not insured via family plan. About 90% of the population in Germany is insured via statutory health insurance; most of the remaining 10% are privately insured. Only certain groups of persons have the option of a private health insurance, such as freelancers, public officers, and employees with an income above an income threshold for compulsory insurance.[7] The present study has access to data from all the statutory health insured persons in Germany. In Germany it is estimated that around 10 million people have no public health insurance.

6.4 Outcomes relevant for Health Technology Assessment

Primarily, we will investigate the value of RWD from European national healthcare registers and statutory health insurance data (claims data) in generating high-quality, accessible, population-based information on breast cancer and ALS (diagnosis, treatment, outcome). Second, we will investigate the application of historical control arms of RCT neglected populations (breast cancer) and synthetic data (ALS) in the attempts to improve external validity, and statistical power and precision (cf. section 5.6). On a higher level the methodology packages of high quality produced in the overall Real4Reg study may be potentially applied for regulatory purposes and HTA.

7. Bias

Any risk of bias related to how data is captured and reported will be described.

8. Effect measure modification

No effect modification will directly be investigated.

Importantly, we will examine heterogeneity within the dataset, in the context in which the data are captured (differences in national healthcare registers and claims data) including coding practices in the four partners, accessibility, representativeness, temporal variation in variables, bias, completeness (missing data), cf. section 2.2.1. But this is unrelated to effect modification.

9. Data sources

9.1 Meta-data about data sources and available software

Four European countries (namely Denmark, Finland, Germany, Portugal, all recognised to have excellent data linkage resources at the national level) will provide RWD from national healthcare registers and claims data.

All available secondary data (exposure, outcomes, covariates), stratified by country are specified in Table A. Table A records the calendar time range used to ascertain cohort entry (index date), as well as the calendar time range of data available for pre-index assessment windows and post-index follow up (study period). The data source name and version are identified, as well as any sampling criteria applied and software applied.

9.2 Coding systems

Coding systems applied for cohort definition, covariates, exposures and outcomes are respectively provided in Table B, Table E and Table G.

More specifically, NOMESCO, ICD10/ICD-10-CM/ICD-10-PCS codes [diagnosis/procedure], ICD-O-3 morphology codes [pathology], VNR [medicine brand names, active ingredient, strength of the medicine, package size], ATC code [medicines] and visit date/visit type [secondary care, in/out/ER patient visits] classifications will be applied.

9.3 Linkage method between data sources

Linkage between data sets will be performed initially on a country basis and are described in Table A. In Denmark a unique pseudonymized person identifier (CPR) is available for each individual living in Denmark, and this will be used to link data. In Finland a similar pseudonymized identification number will be used. No linkage will occur in Germany as only one data source is applied. In Portugal, the linkage between data sets will be at the level of a Master Patient Index (through the *User Number*, available for all citizens).

10. Analysis Plan

10.1 Statistical methods

All table shells and variable definitions are included in section 18 (pages 28-35).

The following steps will be followed:

Use Case 1

1. Pre-processing quality control: Data are inspected for missing values, coding errors and erratic dates. Meta-data will be mapped to the OMOP common data model. We expect the amount of missing data to be minor as the completeness of the data sources and specific field we apply in this study question is close to 100% based on our experience.
2. The population as well as rates of occurrence and outcome will be determined. We will apply the standardized STROBE (Strengthening the Reporting of Observational studies in Epidemiology) standards for reporting non-randomised studies.
3. We will examine and report heterogeneity within the dataset, in the context in which the data are captured (between country differences in healthcare registers and claims data) including:
    i. Coding practices in the four partners
    ii. Accessibility
    iii. Representativeness
    iv. Temporal variation in variables
    v. Bias
    vi. Completeness (missing data)

Use Case 2

4. An algorithm to automatically extract and subset the given RWD data by user defined inclusion/ exclusion criteria and subsequently visualise specific disease trajectories based on data made available in Use Case 1 will be developed. The workflow will allow users to select patients in real-world datasets by demographic characteristics, diagnoses, and medication(s). Following patient selection, the user of the workflow will have the possibility to display summary statistics of the selected patients in comparison to the overall original population to which the in- and exclusion criteria had been applied (4.2 and 4.3), e.g., with respect to the frequency of co-morbidities. The workflow will also allow to display disease trajectories, e.g., via Sankey diagrams
5. Methods to construct synthetic and external control arms for RWD ALS patients will be based on data from the ProACT database: https://ncri1.partners.org/ProACT. To allow for the construction of external control arms, we will implement different propensity score matching algorithms from the literature. This will allow users to identify patients in the external control arm that statistically match the characteristics of patients included in a study provided by the user.
6. Utilising vector space embedding our partners, CSC – IT CENTER FOR SCIENCECS will identify potential changes in prescription policy in a subset of patients.
7. Experimentation and potential implementation of heuristical AI approaches will potentially be included.
8. Depending on the availability of according treatments in the data, we will also explore, whether one of the RCTs contained in ProACT can be emulated using RWD.

10.2 Study size

10.2.1 Breast cancer

Breast cancer is a prevalent condition.

We expect the following study size reported according to participating country.

Denmark: The age standardized annual incidence rate (Nordic) of breast cancer in Denmark is 145.0 per 100,000 females and 1.5 per 100,000 males based on the period 2016-2020.[4] With a population of 5.9 million (female/males: 2.95 million/2.98 million) we expect 4,323 incident primary breast cancer cases per year of intake (breast cancer$_{female/male}$: 4,278/45). From 2000 to 2021 this equates to 51,876 incident breast cancer cases in total (breast cancer$_{female/male}$: 51,336/540).

Finland: The age standardized annual incidence rate (Nordic) of breast cancer in Finland is 145.1 per 100,000 females and 1.0 per 100,000 males based on the period 2016-2020.[4] With a population of 5.6

million (female/males: 2.86 million/2.76 million) we expect 4178 incident primary breast cancer cases per year of intake (breast cancer$_{female/male}$: 4,150/28). From 2000 to 2021 this equates to 50,136 incident primary breast cancer cases in total (breast cancer$_{female/male}$: 49,800/336).

When also including prevalent cases there are 91,279 breast cancer patients.

Germany: The age standardized annual incidence rate of breast cancer in Germany is 112.6 per 100,000 females and 1.1 per 100,000 males based on the year 2018.[5] Within a population of 72.8 million publicly ensured people (females/males: 36.9 million/35.9 million) we expect 41,974 incident primary breast cancer cases per year of Intake (breast cancer$_{female/male}$: 41,579/395). From 2012 to 2021 this equates to 419,740 incident primary breast cancer cases in total (breast cancer$_{female/male}$: 415,790/3,950).

Portugal: The age standardized annual incidence rate of breast cancer in Portugal was 156.0 per 100,000 females and 1.7 per 100,000 males during 2019.[6] With a resident population of 10.3 million (female/males: 5.44 million/4.86 million) we expect 8,529 incident primary breast cancer cases per year of intake (breast cancer$_{female/male}$: 8480/49). From 2000 to 2021 this equates to 187,630 incident primary breast cancer cases in total (breast cancer$_{female/male}$: 186561/1069).

Total expected study size for incident breast cancer patients: 51876 + 50136 + 419740 + 187630 = 709,382

Total expected study size of controls (estimated to be 10-fold greater): 7,093,820

Total expected study size = 7,803,202

10.2.1 ALS

ALS is a rare disease with an annual incidence of 2 per 100,000 persons.

Within the consortium the total population is 94.6 million persons (Denmark, 5.9 million, Finland 5.6 million, Germany 72.8 million, Portugal 10.3 million). Thus, we expect 1900 incident cases of ALS per intake year. All countries include patients from 2008 onwards, whilst Denmark, Finland and Portugal also intake from 2000 to 2007.

This would provide a minimum of (14 years (2008-2021, all countries) * 1892 incident ALS cases/year = 26,488 incident ALS cases).

Additionally, including years 2000 to 2007 (Denmark, Finland, Portugal, total population of 21.8 million persons) would enable inclusion of 3,488 additional incident ALS cases (8 years (Denmark, Finland, Portugal (from 2000-07) * 436 incident ALS cases/year).

Total expected study size of ALS patients: 26,488 + 3,488 = 29,976

Total expected study size of controls (matched 1:10) = 299,760

Total expected study size = 329,736

10.3 Analytic control of confounding and outcome misclassification

Not applicable in the context of this data modelling WP, Use Case 1 and 2.

11. Data management and quality control

To ensure legal compliance and data privacy preservation, each of the data sources will be accessed in a data privacy preserving manner, and each partner will be responsible for their own data. The Real4Reg will generally follow the paradigm of bringing algorithms to the data rather than the other way around.

11.1 Data storage

In Denmark data will be stored on the secure server *Forvaltningsmaskine* of the Danish Health Data Authority. Access is restricted to persons with permission granted by the Danish Health Data Authority and is controlled using a two-step authorisation process.

In Finland the data are stored in the audited remote use environment *Kapseli* provided by the National Health and Social Data Permit Authority Findata. Access is restricted to persons with permission to use granted by the Findata and controlled by a two-step authorisation process.

In Germany, data will be stored in the Health Data Lab (FDZ) of the Federal Institute for Drugs and Medical Devices (BfArM). The developed algorithms will run on the internal data base and only the results will be made available to the researchers.

In Portugal, data will be stored on a server in *Infarmed* and access will be permitted using an authorisation process.

11.2 Independent review of study results

Study results will be available for independent review on the Real4Reg website https://www.real4reg.eu/ .

11.3 Data sharing

Access to the data is restricted as mentioned in section 11.1. in order to comply with data protection regulations, also see section 13. For external review of analyses, access can be applied for from the data authorities granting the permission (Table A).

11.4 Quality control

Pre-processing quality control: Data are inspected for missing values, coding errors and erratic dates. We expect the amount of missing data to be minor as the completeness of the data sources and specific field

we apply in this study question is close to 100% based on our experience. Meta-data will be mapped to the OMOP common data model. The OMOP mapping is checked and validated, e.g. with the OHDSI Data Quality Dashboard tool (https://github.com/OHDSI/DataQualityDashboard).

12. General limitations

Causal inference is a major limitation when analysing RWD. In addition, RWD heterogeneity from different partners is as a limitation. Concerns arise regarding confounding, coding biases, missing or misclassified variables, and especially regarding precise definition of exposure and outcome measurement. In WP1 we attempt to illustrate that RWD use in pre-authorisation and evaluation steps of a medical product is an important source of information – especially when an RCT is not possible for practical or ethical reasons.
General overall limitations of WP1 include problems with data sharing options and/or low prediction performance of the AI/ML-algorithms.

13. Ethical/data protection issues

Real4Reg is entirely registry based and most of the data sources used in this study are currently already used for pharmaco-epidemiological research. The Real4Reg partners from different EU member states will process personal data from individuals which are collected in national/regional electronic health record databases. Due to the sensitive nature of this personal medical data, we strive to take all reasonable measures to ensure compliance with ethical and regulatory issues on privacy. When required, the study protocols will be reviewed by the national data permit authorities (e.g. Findata) and by Institutional Review Boards (IRBs) of the respective participant institutions and/or data sources.

The pseudonymized patient-level data will not leave any of the data holding organisations. Instead developed algorithms will be brought to the data: Each data holding organisation will set up a dedicated server within in a demilitarised zone, where algorithms can be developed, and calculations can take place. Separate data processing agreements will make sure that neither models nor data can leave these servers, hence providing strong data protection. The intended users of the AI algorithms are statisticians in regulatory agencies or universities. They will be informed about their interaction with AI techniques. Users will be appropriately trained to understand the capabilities, limitations and risks of AI algorithms.

The consortium asserts that all procedures contributing to this research comply with the ethical standards of the relevant national laws of all participating countries and according to the Helsinki Declaration.

All consortium partners have a well-developed mechanism to ensure that European and/or local regulations dealing with ethical use of the data and adequate privacy control are adhered to. All data sources will be processed in compliance with relevant legislation and guidance and in line with the General Data Protection Regulation (EU 2016/679). Specifically, Denmark, Finland and Portugal have obtained local approval for their contributions from the respective local Data Protection Agencies. For Germany no direct access to personal data will occur, therefore no approval from a state agency is necessary.

We will statistically attempt to assess whether predictions made by our models are unintentionally impacted by ethnicity and gender. If we find such biases, we will try to eliminate them (e.g. by removing according variables in the training data). If this is unsuccessful, we will raise an according warning.

There are no further ethical risks as models will only be used to support regulatory decision-making, but not replace it. More specifically, our models will provide additional sources of evidence to the regulator, which he/she can consider jointly with the clinical trial data provided by companies.

According to European law, registry and claims data can be used for research without obtaining individual informed consent.

14. Amendments and deviations

This section will document amendments and deviations. Country-specific adjustments of analyses according to data availability and national differences in coding practices may be performed.

15. Plans for communication of study results

Overall, Real4Reg is committed to a rapid and effective dissemination, exploitation and communication of project results as well as newly generated knowledge to all relevant audiences. This includes the medical, pharmacist and applied regulatory science community as well as all other health care professionals and public health experts including health insurances, regulators, HTA and policy makers. Another important target audience will be the general public as the successful implementation and extension of the effective use of RWE in the regulatory and HTA context will require active participation of all patients receiving drug treatments. A dedicated work package WP6 will handle all dissemination tasks.

In brief, the Real4Reg project will:

1. Be promoted online via a public website. This project website will contain information about the overall scope of the project and background, as well as information on individual work packages, the

project team, events, and results. It will also include a section aimed at patients and the general public.

2. Include press releases and a newsletter to raise public awareness of the project as well as social media accounts from the partner organisations (LinkedIn and *X*).
3. Be regularly presented at international medical and scientific conferences and will be published in well-known peer-reviewed national and international scientific peer-reviewed journals. The Real4Reg consortium embraces the concept of providing open-access whenever possible for a timely dissemination within the scientific and regulatory/ HTA community.
4. Organise events such as workshops and symposium, to publicise the results and their implications for society, including activities dedicated for patients.

16. Timeline

The project is a four-year project performed during the period January 1, 2023 to December 31, 2026.

Tentativis specific deliverables/milestones related to WP1, Use Case 1 and 2, applied to breast cancer include:

Month	Deliverable/milestone	Date/month
Consortium meeting	M	January and June 2023
Data access	M	September 2023
Reports of scientific results (M24 and M40)	D	December 2024 and April 2026
Launch of website	D	June 2023
Data management plan (received from partner)	D	June 2023
First version of common data model (CDM)	M	June 2023
First version of web-based information portal	M	June 2023
Registration in the EU PAS Register	D	June 2023
Participation in RWD workshops	M	Annually (in September)
First version of common data model (CDM)	D	December 2024
First version of web-based information portal	D	December 2023
Interim report		June 2025

17. References

1. de Jong J, Emon MA, Wu P et al. Deep learning for clustering of multivariate clinical patient trajectories with missing values. 2019 *Giga Science*, 8 (11), 1-14, doi: https://doi.org/10.1093/gigascience/giz134

2. Lentzen M, Linden T, Veeranki S et al. A Transformer-Based Model Trained on Large Scale Claims Data for Prediction of Severe COVID-19 Disease Progression. 2022 *BMJ Yale*. doi: https://doi.org/10.1101/2022.11.29.22282632

3. Birkenbihl C, Ahmad A, Massat N et al. Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories. 2023, *Sci Rep* 13:2897 (2023). doi: https://doi.org/10.1038/s41598-023-30038-8

4. Nordcan https://gco.iarc.fr/media/nordcan/factsheets/92/en/countries/208/breast-180-denmark-208.pdf (last accessed May 2, 2023)

5. RKI https://edoc.rki.de/bitstream/handle/176904/8320/cancer_germany_2015_2016_2.pdf (last accessed May 2, 2023)

6. RON https://ron.min-saude.pt/media/2214/ron-2019_new_v8f.pdf (last accessed May 2, 2023)

7. https://www.bundesgesundheitsministerium.de/private-krankenversicherung

18. Tables

Table A. Meta-data about data source and software

Table A1 Denmark (Note: First row in Table A1 is relevant for tables A2, A3, A4)

Data source(s)	Description	Study period	Eligible cohort entry point	Data extraction date/version	Applied to study populations	Data sampling/ extraction criteria	Data linkage	Type(s) of data	Data conversion	Software to create study population
Civil Registration System (CPR)	Contains individual-level information on personal-identifiable number.	2000-2022	1995-2022 (BC); 1995-2022 (ALS)	Presumably July 2023	Both BC and ALS use	Patients ≥18 years	CPR	Registry	OMOP	R/SAS/Python
National Prescription Register (LRS)	Contains information on all sales of human and veterinary medicinal products in	2000-2022	1995-2022 (BC); 1995-2022 (ALS)		Both BC and ALS use		CPR	Registry		
National Patient Register (LPR)	Contains information on all hospital contacts.	2000-2022	1995-2022 (BC); 1995-2022 (ALS)		Both BC and ALS use		CPR	Registry		
Hospital register	Contains information on medication use in public hospitals (currently under	2000-2022	2018-2022 (BC); 2018-2022 (ALS)		Both BC and ALS use		CPR	Registry		
Cancer register	Contains information on all cancer diagnoses.	2000-2022	1995-2022 (BC)		Only BC use		CPR	Registry		
Laboratoriedatabasen(LAB)	Biopsy data	2000-2022	2008-2022 (BC)		Only BC use		CPR	Registry		
Sygesikringsregisteret (SSR)	1st trimester GP visit	2000-2022	1995-2022 (BC)		Only BC use		CPR	Registry		
Sygesikringsregisteret (SSR2)	1st trimester GP visit	2000-2022	2014-2022 (BC)		Only BC use		CPR	Registry		
Psychiatric Research Registry (PCR/PSYK)	Information on all hospital contacts with psychiatric	2000-2022	1995-2022 (BC)		Only BC use		CPR	Registry		
Causes of death register	Death dates, causes, how the cause was ascertained	2000-2022	2018-2022 (BC); 2018-2022 (ALS)		Both BC and ALS use					
National Pathology Registry (PATH)	SNOMED codes, hormone receptor status	2000-2022	2018-2022 (BC); 2018-2022 (ALS)		Both BC and ALS use		CPR	Registry		
Medical Birth Registry (MFR)	Parity, births, complications, BW, BL	2000-2022	2018-2022 (BC); 2018-2022 (ALS)		Only BC use		CPR	Registry		

Table A2 Finland

Care register for health care	Hospitalisations	2000-2022	1995-2022	June 2023	specific dates in relation to index date (pref 1995 onwards to cover just ICD-10)		by pseudonymised personal identification number			R/Python
	Specialised healthcare outpatient visits	2000-2022	1998-2022							
	General healthcare outpatient visits	2000-2022	2011-2022							
	(diagnoses, procedures, required level of assistance at discharge. Also medications + vaccines in the newer data since ca. 2015 but the completeness of these med/vacc data has not been assessed yet	2000-2022	2015-2022							
Cancer register	Confirmed cancer cases, tumour type, morphology,	2000-2022 2000-2022	1995-2022		only BC use case					
Special reimbursements	comorbidity information, reimbursement code &	2000-2022 2000-2022	1995-2022							
Kanta physiological measurements	measurement type, results, reference values, units…)	2000-2022 2000-2022	2014-2022							
Kanta laboratory results	measurement type, results, reference values, units…)	2000-2022	2014-2022			confirmed results only				
Dispensed prescriptions, Prescription	ATC, drug name, purchase date, amount, strength, dosing in newer data…	2000-2022 2000-2022	1995-2022							
Kanta electronic prescription database		2000-2022 2000-2022	2010-2022							
Causes of death register	death dates, causes, how the cause was ascertained	2000-2022 2000-2022	1995-2022							
Statistics Finland socioeconomic	education, occupation, income/wealth, family size	2000-2022 2000-2022	1995-2022							
Sickness benefit information	ICD-10, duration, allowance type, occupation information. amount paid…	2000-2022 2000-2022 2000-2022	1995-2022		Possible to use in ALS use cases					

Table A3 Germany

Health Data Lab	Contains individual-level information with a personal-identifiable number on all people covered by statutory health insurances (SHI). Information comprises demographics, insurance status and days covered, outpatient medicinal products prescriptions, inpatient and outpatient diagnoses and procedures, further health care sector information (e.g., care status, remedies and aids)	2000-2022	Presumably 2008-2022	Presumably 01.06.2023 Possibly preliminary data only Version 1.0	BC and ALS	All patients 18 years and above insured by a public health care insurance	Not applicable	Claims data of public health insurance providers	OMOP	R/Python

Table A4 Portugal

National User Register (RNU)	Contains individual-level information on personal-identifiable number, sex, birth date, residence	2000-2022	1995-2022 (BC); 1995-2022 (ALS)	Presumably July 2023	Both BC and ALS use	Patients ≥18 years	RNU	Registry	OMOP	R/phyton
National Electronic Reimbursed Dispensing Register (CCM-SNS)	Contains information on all reimbursed sales of human products in Portugal	2000-2022	2012-2022 (BC); 2012-2022 (ALS)		Possible use for BC - History of hormone replacement therapy use		RNU	Claims Data		
Hospital Morbidity Database (BDMH)	Inpatient and outpatient events and procedures (ICD-10-CM/PCS) in public hospitals	2000-2022	2000-2022 (BC); 2000-2022 (ALS)		Both BC and ALS use		RNU	Registry		
Primary care databse (BICSP)	Outpatient events and procedures (ICPC-2)	2000-2022	2000-2022 (BC); 2000-2022 (ALS)		Both BC and ALS use		RNU	Registry		
Cancer register database (RON)	Contains information on all cancer diagnoses, histology and procedures (ICD Oncology 3rd Ed)	2000-2022	2000-2022 (BC)		Only BC use		RNU	Registry		
Death register database (SICO)	Contains information on death dates, causes (ICD-10-CM)	2000-2022	2014-2022 (BC); 2014-2022 (ALS)		Both BC and ALS use		RNU	Registry		

Table B. Cohort entry defining criterion

Table B1 Denmark (Note: First row in Table B1 is relevant for tables B2, B3, B4)

Study population name(s)	Day 0 Description	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type	Diagnosis position[2]	Incident with respect to…	Pre-specified	Varied for sensitivity	Source of algorithm
BC	Primary incident BC diagnoses using the WHO International Classification of Diseases, version 10 converted (ICD-10 converted) codes from 2000-2003 and ICD-10 codes from 2004 onwards.	Incident diagnosis of primary breast cancer	Single	[-5 years;0[Incident, no prior other malignancy (except for carcinoma in situ of the cervix or basal cell carcinoma or squamous cell carcinoma of the skin that has been previously treated with curative intent), without being disease-free for more than 5	Cancer Registry (IP, OP)	ICD10	any	breast cancer diagnosis	no	no	specifically developed
ALS	First diagnosis date between 2000-2021. We will identify incident ALS diagnoses using the version 10 (ICD-10) codes from 2000 onwards	Incident diagnosis of ALS	Single	[-5 years;0[(5-years)	Hospital (IP, OP)	ICD10, ATC	any	ALS diagnosis	no	no	specifically developed

Table B2 Finland

BC	First diagnosis date between 2000-2021	Incident diagnosis of primary breast cancer		can be added if we want to restrict to incident (not reoccurring cases)	IP/OP (cancer registry)			breast cancer diagnosis		no	specifically developed
ALS	First diagnosis date between 2000-2021	Incident diagnosis of ALS		5 year	IP/OP (specialised healthcare)			ALS diagnosis	yes	x (restricted to main diagnosis)	specifically developed

Table B3 Germany

BC	Index date will be the date of the first (incident) ICD-10 diagnosis after a gapless SHI-covered period of at least 5 years.	First diagnosis of primary breast cancer	Single	incident	[-5 years;0[	ICD10	any	breast cancer diagnosis	no	no	specifically developed
ALS	Index date will be the date of the first (incident) ICD-10 diagnosis and Riluzole dispensing after a gapless SHI-covered period of at least 2 years.	First diagnosis of ALS	Single	incident	[-2 years; 0[	ICD10, ATC	any	ALS diagnosis	no	no	specifically developed

Table B4 Portugal

BC	First diagnosis date between 2000-2021, according the European Network of Cancer Registries (ENCR) https://www.researchgate.net/publication/359230862_ENCR_Recommendation_CODING_INCIDENCE_DATE	First diagnosis of primary breast cancer	Single	5 year	IP, OP	ICD-10 CM, ICD Oncology 3	any	breast cancer diagnosis	no	no	specifically developed
ALS	First diagnosis date between 2000-2021	First diagnosis of ALS	Single	2 year	IP, OP	ICD-10 CM, ATC	any	ALS diagnosis	no	no	specifically developed

Table C. Inclusion Criteria

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type	Diagnosis position[2]	Applied to study populations	Pre-specified	Varied for sensitivity	Source for algorithm
≥18 years		Before selection of index date	na	na	na	any	BC, ALS	Yes	No	
Observable patient time	Without bridging of gaps	Before selection of index date	na	na	na	any	BC, ALS	Yes	No	specifically developed

Table D. Exclusion Criteria

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type	Diagnosis position[2]	Applied to study populations	Pre-specified	Varied for sensitivity	Source for algorithm
No records of treatment	For German claims data, single years without health reco	Before selection of index date	Whole study	IP, OP	ICD/ATC	NA	ALS	Yes	No	specifically developed
"Washout" period/Primary cancer		After selection of index date	[- 5 years; 0[	Dependent on country (IP, OP/ NA)	ICD 10 / ICD 8 (/ ICD7)	Any	BC, ALS	Yes	No	specifically developed
Country specific washout period/ALS		After selection of index date	at least [2 years; 0[for ALS	Dependent on country (IP, OP/ NA)	ICD 10 / ICD 8 (/ ICD7)	Any	BC, ALS	Yes	No	specifically developed

Table E. Predefined Covariates

Table E1 Denmark (Note: First row in Table E1 is relevant for tables E2, E3, E4)

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type	Diagnosis position[2]	Applied to study populations:	Pre-specified	Varied for sensitivity	Source for algorithm
Pregnancy	MFR	Categorical births/still	[study entry; 0[	Patient Registry	Registry	any	BC only	yes	no	specifically developed
		Parity	[study entry; 0[	Patient Registry	Registry	any	BC only	yes	no	
		Complications	[study entry; 0[	Patient Registry	Registry	any	BC only	yes	no	
		Birth outcomes	[study entry; 0[	Patient Registry	Registry	any	BC only	yes	no	
	Sygesikringsregisteret	First trimester GP visit	[study entry; 0[	Patient Registry	Registry	any	BC only	yes	no	
Age	CPR		[study entry; 0[	Administrative registry	Registry	any	BC and ALS	yes	no	
Sex	CPR		[study entry; 0[	Administrative registry	Registry	any	BC and ALS	yes	no	
SES	CPR	Address	[study entry; 0[	Administrative registry	Registry	any	BC and ALS	yes	no	
OC use	Prescription registry	ATC code	Time of diagnosis/during FU	OP	Registry	any	BC only	yes	no	
HRT use	Prescription registry	ATC code	Time of diagnosis/during FU	OP	Registry	any	BC only	yes	no	
Co-morbidities	LPR	ICD codes	Time of diagnosis/during FU	OP, IP	Registry	any	BC and ALS	yes	no	

Table E2 Finland

Pregnancy	inpatiet/outpatient	ICD-10 and NOMESCO	Time of diagnosis/during follow-	In- and outpatient	Register	any	BC only	Yes	No	specifically developed
Age	CPR	continuous	[study entry; 0[	Administrative registry	Register	any	BC and ALS	Yes	No	
Sex	CPR	binary	[study entry; 0[	Administrative registry	Register	any	BC and ALS	Yes	No	
OC use	Prescription register	ATC code	Time of diagnosis/during follow-up	OP	Register	any	BC only	Yes	No	
HRT use	Prescription register	ATC code	Time of diagnosis/during follow-up	OP	Register	any	BC only	Yes	No	
Comorbidites	Inpatient/outpatient healthcare registers, Prescription register, Special reimbursement register	ICD10, ATC, special reimbursement codes	Time of diagnosis/during follow-up	IP, OP	Register	any	BC and ALS	Yes	No	

Table E3 Germany

Pregnancy	Claims data of public health insurance providers	ICD10 GM code	[study entry; 0[	In- and outpatient data	Claims data	any	BC only	Yes	No	specifically developed
Age	Claims data of public health insurance providers		[study entry; 0[				BC and ALS	Yes	No	
Sex	Claims data of public health insurance providers		[study entry; 0[							
OC use	Claims data of public health insurance providers	ATC code	Time of diagnosis/during follow-up				BC only	Yes	No	
HRT use	Claims data of public health insurance providers	ATC code	Time of diagnosis/during follow-up				BC only	Yes	No	
Comorbidites	Claims data of public health insurance providers	ICD10 GM code	[study entry; 0[	In- and outpatient data		any	BC and ALS	Yes	No	

Table E4 Portugal

Pregnancy (Availability to be confirmed)	BICSP, BDMH	ICD-10 CM, ICPC-2 codes	[study entry; 0[	IP, OP	Registry	any	BC	yes	no	specifically developed
Age	RNU		[study entry; 0[	IP, OP	Registry	any	BC, ALS	yes	no	
Sex	RNU		[study entry; 0[	IP, OP	Registry	any	BC, ALS	yes	no	
OC use	CCM-SNS	ATC code	[study entry; 0[	OP	Registry	any	BC	yes	no	
HRT use	CCM-SNS	ATC code	[study entry; 0[	OP	Registry	any	BC	yes	no	
Comorbidites	BICSP, BDMH	ICD-10 CM, ICPC-2 codes	[study entry; 0[	IP, OP	Registry	any	BC, ALS	yes	no	

Table G. Exposure and Outcome

Exposure and Outcome name	Outcome measurement characteristics	Primary outcome?	Type of outcome	Washout window	Care Settings[1]	Code Category	Diagnosis position[2]	Applied to study populations:	Pre-specified	Varied for sensitivity	Source of algorithm
Incidence	Desriptive		Derived		Hospital		-		yes		
Prevalence	Desriptive		Derived		Hospital		-		yes		
Survival time			Derived				-		yes		
All BC - Overall survival (OS) as a time-dependent variable					Hospital				yes		
Metastatic BC - Progression-free survival	Rate				Hospital				yes		
Non-metastatic - Disease-free survival (DFS) as a time-dependent variable					Hospital				yes		
5-year and 10-year disease free survival			Derived		Hospital				yes		
Non-metastatic - disease-free survival (ie time with no metastases)	Rate				Hospital				yes		
5 and 10-year Mortality rate (overall and BC specific mortality rate)	Rate				Hospital				yes		
Signals for disease progression					Hospital				yes		
Metastasis					Hospital				yes		
Invasion					Hospital				yes		
Stage					Hospital				yes		
TMN					Hospital				yes		
Morbidities			ICD-8/10		Hospital				yes		
Secondary cancers	Signals		ICD-10		Hospital				yes		
Death		yes	Status code in CPR or ICD cause of death code in Causes of death registry	-	-	-	-		yes		
Signals for disease treatment					Hospital				yes		
Radiotherapy			Procedure codes		Hospital				yes		
Chemotherapy	Signals		Procedure codes		Hospital				yes		
HRT use			ATC codes		Prescription				yes		
Oophorectomy			Procedure codes		Hospital				yes		
Mastectomy			Procedure codes		Hospital				yes		

19. List of Appendices

Table appendix 1. Analysis specifications

	Primary	**Secondary 1**	**Secondary 2**
Hypothesis:	Explorative -> we can provide additional value for pharmaceutical approval processes		
Study population(s)	BC, ALS		
Outcome:	Death, comorbidity, signals of disease/treatment progression		
Software:	R, Python, SAS		
Model(s):			
Confounding adjustment method			
Bivariate			
Multivariable			
Other			
(specify details)			
Missing data methods			
Missing indicators			
Complete case			
Last value carried forward			
Multiple imputation (specify variables)			
Other (please specify)			
Subgroup Analysis			

HORIZON-HLTH-2022-TOOL-11

Real4Reg: Development, optimisation and implementation of artificial intelligence methods for real world data analyses in regulatory decision-making and health technology assessment along the product lifecycle

Title	Work Package (WP) 2, Use Case 3 **Using Real World Data to identify adverse drug reactions and impact of regulatory interventions on prescriptions – oral fluoroquinolones as the use case**
Protocol version identifier	1.1
Date of last version of protocol	28-03-24
EU PAS register number	EUPAS105544
Active substance	Fluoroquinolones (ATC J01MA)
Medicinal product	Not specified
Product reference	Not specified
Procedure number	Not specified
Marketing authorisation holder(s)	Not specified
Joint PASS	No
Research question and objectives	The overall objective is the preparation of a good practice example for post authorisation safety studies (PASS) based on Real-World Data (RWD); with the specific aim of improving methods for estimation in observational data. We assess how RWD can be used to generate high-quality, population-based information on the risk of prespecified adverse drug reactions (ADRs), and to evaluate the impact of regulatory warnings on the use of broad-spectrum antibiotics by using fluoroquinolones (FQ) and risk of prespecified ADRs in adults as the use case. The specific objectives are to: 1. examine whether there were changes in antibiotic prescribing and patient characteristics following the regulatory interventions recently established on FQ. 2. examine the estimated risk of ADRs in patients with FQ prescription retrievals, including characteristics before and after FQs authorization changes. 3. describe similarities and differences (heterogeneity) between the available data sources from four participating countries to examine whether and how the heterogeneity in data leads to heterogeneity in results 4. explore whether ADRs can be predicted on the individual patient level using Artificial Intelligence (AI) / Machine Learning (ML) techniques
Countries of study	Denmark, Finland, Germany, Portugal
Author	Anna-Maija Tolppanen, Sirpa Hartikainen, Anne Paakinaho

Members of Work Packages 1-3 of the Real4Reg project and associates contributed to the protocol and adopt it:	Aborageh, Mohamed Adewuyi, Davis Arzideh, Roxana Becker, Cornelia Bräuner, Elvira Braithwaite, Billy Costa, Inês Ehrenstein, Vera Fernandes, Joana Froehlich, Holger Furtado, Cláudia Haenisch, Britta Hartikainen, Sirpa Heß, Steffen Horváth-Puhó, Erzsébet Kallio, Aleksi Korcinska Handest, Monika Roberta Nagy, Dávid Paakinaho, Anne Peltner, Jonas Pfeifer, Kerstin Pylkkanen, Liisa Rajamaki, Blair Roethlein, Christoph Russek, Martin Schneider, Katharina Silva, Célia Tolppanen, Anna-Maija Vancraeyenest, Aurélie Vo, Thuan Wicherski, Julia

Table of Contents for WP2, Use Case 3 Fluoroquinolones

1. List of abbreviations

ACNU Active Comparator New User

ADR adverse drug reaction

AI Artificial Intelligence

ATC Anatomical Therapeutic Chemical

CDM common data model

CI confidence interval

FQ fluoroquinolones

HTA Health Technology Assessment

ICD international classification of diseases

IPTW inverse probability of treatment weighting

IRB Institutional Review Board

ITT intention-to-treat

ML Machine Learning

NNH number needed to harm

OMOP Observational Medical Outcomes Partnership

PASS post authorisation safety studies

QC quality control

RWD Real-World Data

WP work package

2. Research question

The overall objective is the preparation of a good practice example for post authorisation safety studies (PASS) based on Real-World Data (RWD); with the specific aim of improving methods for risk estimation in observational data. This protocol describes the analyses for use case 3 (fluoroquinolones and risk of prespecified adverse drug reactions [ADRs] in adults) of the Real4Reg project. The target population of this study are adults who receive a prescription of an oral antibiotic (later referred to as new users of antibiotics in this protocol).

2.1 Study objectives

To evaluate how RWD can be used to generate high-quality, population-based information on the risk of prespecified ADRs, and to evaluate the impact of regulatory warnings on the use of broad-spectrum antibiotics prescriptions.

The specific objectives are to:

1. Examine whether there were changes in prescription patterns of broad-spectrum antibiotics and patient characteristics due to regulatory interventions recently established on fluoroquinolones (FQ).
2. Examine the estimated risk of prespecified ADRs in patients with FQ prescription retrievals, including characteristics before and after FQs authorization changes.
3. Describe similarities and differences (heterogeneity) between available data sources from four participating countries and to examine whether heterogeneity in data leads to heterogeneity in results and how this should be taken into account in reporting
4. Explore whether ADRs can be predicted on the individual patient level using Artificial Intelligence (AI) / Machine Learning (ML) techniques.

2.2 Specific tasks

1. Data preprocessing
   1.1 Provide data access and carry out data preprocessing tasks (Quality Control (QC) and conversion to the Observational Medical Outcomes Partnership (OMOP)- common data model (CDM))
   1.2 Provide a metadata catalogue and summary
2. Perform a descriptive drug utilization study of FQ and other broad-spectrum antibiotic prescriptions in 2010-2021 in four European countries (time period may depend on the data availability within each participating countries).
   Including:
   - overall time trend
   - impact of safety regulatory interventions on FQ prescription rate and FQ user characteristics
   - assessment of time trends in confounding
3. Estimate the absolute and relative risk of pre-specified ADRs (aortic aneurysm and dissection, cardiac arrhythmia and sudden cardiac death, acute toxic liver diseases, or peripheral polyneuropathy) with active comparator new user (ACNU) design.
   - comparison of common vs site-specific propensity scores

4. Predict drug-related safety issues for individual patients with artificial intelligence (AI)/machine learning (ML)
5. Assess data heterogeneity and whether this leads to heterogeneity of results

3. Study design

The study uses secondary data. Data sources from the participating countries (Denmark, Finland, Germany and Portugal) are listed in Table A. Two study designs will be applied:

3.1 A descriptive cross-sectional drug utilization study that illustrates the changes in prescription retrievals of FQ and other broad-spectrum antibiotics and user characteristics during the study period (2010-2021). We will assess the proportion of FQ prescriptions per all oral antibiotic prescriptions during each year of the study period. Furthermore, where data permits in each partner, the change in age and sex distribution and prevalence of comorbidities in FQ users, and amoxicillin or cephalosporin users over time will be assessed.

3.2 A cohort study with multiple eligibility-based entries per person using an ACNU design to assess the risk of ADRs associated with FQ use compared with other broad-spectrum antibiotics use (cephalosporins and amoxicillin) (Figure 1). The incidence of ADRs during 90- and 365-day follow-up periods among the exposure groups will be calculated and the relative risk increase among FQ users will be estimated using hazard ratios. Finally, risk difference and number needed to harm (NNH) will be calculated.

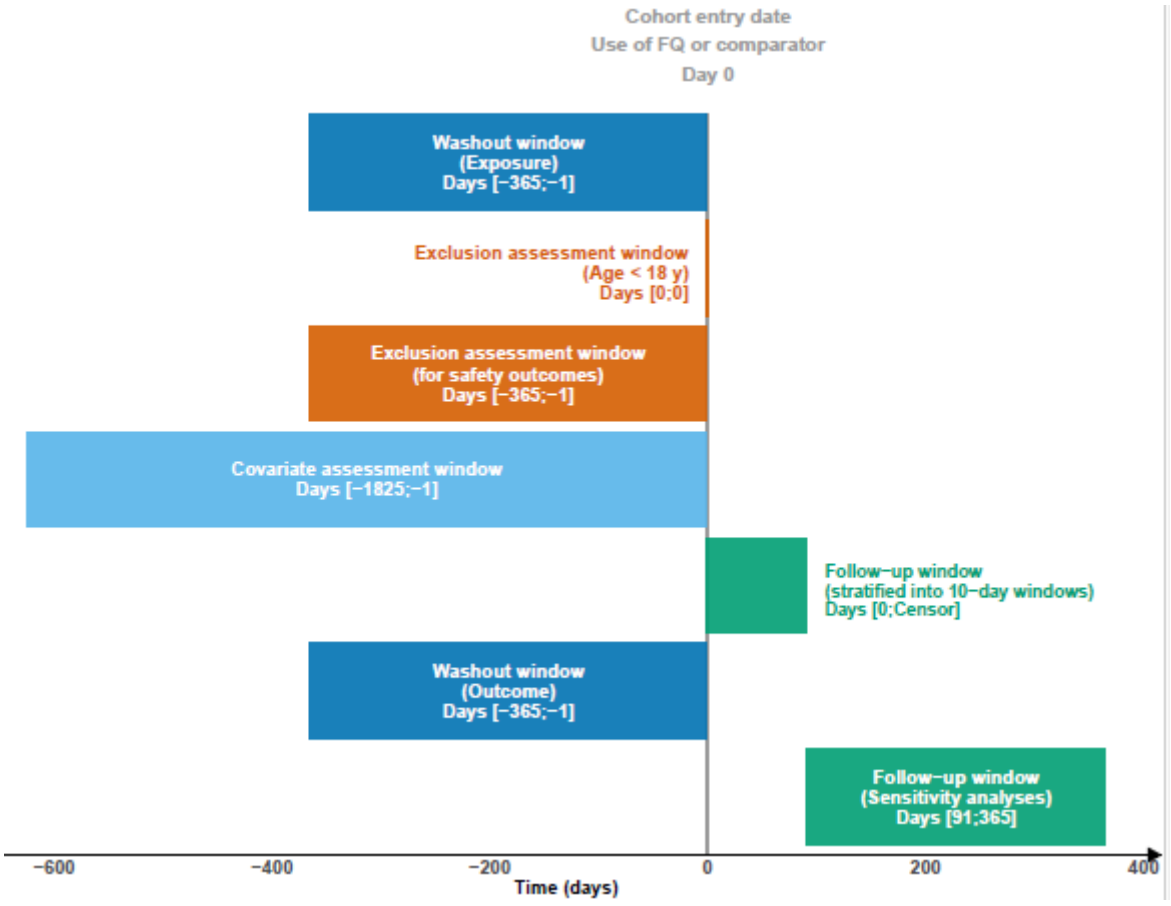

Figure 1. Visualisation of study design in the ADR studies.

4. Source and study populations

The source populations are adult populations (≥ 18 years on the date of prescription retrieval) of Denmark, Finland, Germany, and Portugal. The study participants are persons with written (Germany) or purchased prescription (Denmark, Finland, Germany, Portugal) for oral antibiotics in 2010-2021, identified from national healthcare registers (Denmark, Finland and Portugal) and claims data (Germany). Inclusion and exclusion criteria are described in Tables C and D.

5. Exposure definition and measurement

In the ADR studies, the exposure is defined as new use of oral FQ or comparison broad-spectrum antibiotics using a one-year washout as described in Tables B and C and the index date is defined as the beginning on new FQ/comparison antibiotic use.

Comparison antibiotics for the ADR studies:

1. Aortic aneurysm and dissection, cardiac arrhythmias, sudden cardiac death, and polyneuropathy: For these ADRs, the comparison group contains new users of amoxicillin, amoxicillin with clavulanic acid, or cephalosporins (commonly used with comparable indications as FQs)
2. Acute liver injury due to fluoroquinolones:
   For this outcome comparison group consists of new users of cephalosporins, because amoxicillin is listed to be associated with drug-induced liver injury in the American College of Gastroenterology Guidelines 2021

Included persons are permitted to have multiple antibiotic exposure periods (and multiple index dates), provided that each exposure initiation meets the inclusion criteria and exclusion criteria (Tables C and D). Intention-to-treat (ITT) analyses with fixed 90-day follow-up per prescription are conducted. The follow-up is stratified into 10-day windows. Sensitivity analyses with 365-day follow-up will be performed.

In addition to the 90- and 365-day follow-up we will also use a 2 years time period prior to the first documented prescription of the drug. This time window will be used to train AI/ML models predicting adverse drug reactions.

In addition to the ITT analyses, we will perform sensitivity analyses using a *per protocol* approach in which the follow-up is censored on the date of exposure change (i.e., FQ user initiates comparison antibiotic, or the user of comparison antibiotic initiates FQ use) if this occurs during the 90- or 365-day follow-up. Additional purchases of exposure antibiotic (e.g. repeated FQ prescription in the 90- or 365-day time window) do not affect the censoring. Alternative approaches such as time-dependent exposure modelling are considering if the number of repeated prescriptions is significant.

6. Outcome definition and measurement

Outcomes are listed in Table F. Drug utilization study outcomes (calculated separately for all four countries) are identified from the prescription databases of each country. Dispensed prescriptions are used for Denmark, Finland and Portugal, written prescriptions are used for Germany (see Chapter 9.1). The reference population for each specific year in each country is identified from national data sources. The following outcomes are calculated for the drug utilization study.

- The number of FQ and comparison antibiotics (amoxicillin with or without clavulanic acid or cephalosporins) prescriptions per reference population (persons aged ≥18 years) per year
- The proportion of FQ and comparison antibiotics (amoxicillin with or without clavulanic acid or cephalosporins) prescriptions of all oral antibiotic prescriptions
- The change in user characteristics over the study period

The prespecified ADRs are identified from the hospital discharges and outpatient visits in all countries. In addition, causes of death are used in Denmark, Finland and Portugal. Incident outcomes are identified using a one-year washout period that will be extended if possible (dependent on the available database in each of the four participating countries). The following outcomes are assessed, and a separate study population is built for each outcome

- Sudden cardiac death or cardiac arrhythmia
- Sudden cardiac death separately
- Aortic aneurysm and dissection
- Acute toxic liver disease (as a proxy for drug induced liver injury)
- Polyneuropathy (Drug-induced peripheral polyneuropathy)

Exclusion criteria for specific studies are given in Table D and described in detail in section 9.3.

The validity of the Danish and Finnish healthcare registers has been reviewed previously (Laugesen et al. 2021, Schmidt et al. 2014, 2015, 2019, Sund 2012), and a high positive predictive value for diseases of the circulatory system diagnoses in Finland (Keskimäki and Aro 1991) and cardiovascular diagnoses in Denmark (Sundboll et al. 2016) has been reported. However, it should be noted that the previous validation studies in Finland were performed for inpatient registers and the introduction of specialised healthcare polyclinic visits in 1998 and primary healthcare use since 2011 have likely led to improved validity for conditions that can be treated and diagnosed in outpatient settings, although this likely has more impact on capturing covariates than the outcomes in this study.

7. Bias

In the ADR studies, confounding by indication will be controlled by ACNU design. In addition, we will use inverse probability of treatment weighting (IPTW) to account for confounding. The weights are derived from predefined covariates (Table E2). In addition, an ML approach is used to derive the weights.

Prevalent user bias is controlled by using a one-year washout to identify new users. We use a one-year washout instead of a longer washout or restriction to first-time initiators to increase the generalizability of the results. Because the data sources of all countries do not cover the hospital-administered drugs, we exclude persons who were hospitalized due to infection during the washout to address the prevalent user bias. However, this may also lead to exclusion of frail persons so sensitivity analyses without this exclusion will also be performed.

As the duration of exposure is short, and well defined, we expect no healthy adherer bias. We expect the misclassification of exposure to be relatively small and nondifferential between the exposure groups. Misclassification of outcome may occur due to differences in coding practices, especially for acute toxic liver diseases (drug induced liver disease) or drug-induced peripheral polyneuropathy. It is possible that there is between-country and within-country variation in differentiation diagnostics for these conditions, which will be considered as an explanation if there is significant variation in the incidence rates of outcomes per exposure group.

The use of the causes of death register for ascertaining the outcomes likely leads to an increased number of events in comparison to using hospital/outpatient diagnosis. The three countries that use causes of death register data have very similar processes for ascertaining the cause of death and registration of them. The European Commission Regulation (EC) No 1338/2008 confirms the variables, specifications, and metadata which the EU Member States have to supply concerning the statistics on causes of death. In Finland, Denmark and Portugal, death certificates are issued by the physician who establishes the death and causes are recorded using the ICD-10 coding. In Finland, most causes of death are based on the clinical data, and autopsied are confirmed in less than 20% of cases (14.6% deaths confirmed with forensic and 3.6% by medical autopsy in 2020). If the information in the death certificate is deficient, inconsistent or difficult to classify, a clarification request is issued.

8. Effect measure modification

We will evaluate whether the associations between FQ and the ADRs are similar across sites by comparing risk ratio and risk difference estimates and their 95% Confidence Intervals (CIs) and evaluate whether they differ in pre- and post-regulatory warning period using the EMA warning date of 5.10.2018 as the cutoff for index dates. Alternative dates, such as European Commission dates for legally binding decisions on restriction of use (14.2.2019, 11.3.2019) are considered if the drug utilisation studies indicate this is necessary. No other effect modification analyses are planned, unless we observe a time trend in covariates in the drug utilisation study.

9. Data sources

Data sources are described in Table A.

9.1 Data sources and coding system for exposure, outcomes and covariates

All four participating countries have excellent data linkage resources at the national level and will provide RWD from national healthcare registers and claims data. Data sources used to ascertain information on exposure, outcomes and covariates are listed in Tables B, E, F.

Exposure to antibiotics is identified from dispensed prescriptions (all for Denmark, reimbursed for Finland and Portugal) and claims data (Germany) using Anatomical Therapeutic Chemical (ATC) codes. The dates of dispensing (Denmark, Finland, Portugal) or prescription writing (Germany) days are used to ascertain the beginning of follow-up. Germany has data on prescribed drugs available until 2018 and on both prescribed and dispensed drugs from 2019 onwards; data on prescribed drugs only are used for Germany to be consistent.

There is variation for settings from which the drug use is captured. The Danish data cover prescriptions purchased from community-dwelling settings and drugs administered in hospitals. The Finnish and Portuguese data sources cover prescriptions purchased in community-dwelling settings and residential care, but not drugs administered in hospitals. The German data cover claims from community-dwelling and long-term care settings.

Outcomes are identified from hospital encounter registers in Finland and Denmark and additionally from primary care data in Finland. In Portugal, disease and treatment registers (including primary care) and claims in Germany are used. In addition, causes of death register is used in all other countries except Germany.

Covariates are identified from the dispensing (ATC codes) codes from the prescription databases and diagnosis (ICD-10, ICPC-2) and procedure (NOMESCO, ICD-10-CM/PCS) codes from the healthcare registers and databases described above.

9.2 Linkage method between data

In Denmark and Finland, the data are linked based on pseudonymised unique person identifiers. In Portugal, the linkage is performed using a Master Patient Index (through the User Number, available for all citizens) followed by a pseudonymised unique person identifier. No linkage will occur in Germany as only one data source is applied.

10. Analysis plan

Country-specific adjustments of analyses according to data availability and national differences in coding practices may be performed. The estimated study size is 88 million, as almost the entire source population of about 88 million are estimated to have an oral antibiotic prescription at some point during the study period.

10.1 Data preprocessing

Data are inspected for missing values, coding errors and erratic dates and converted to the OMOP CDM. Meta-data will be mapped to the OMOP common data model. We expect the amount of missing data to be minor as the completeness of the data sources and specific field we apply in this study question is close to 100% based on our experience.

10.2 Drug utilization study

Outcome: fluoroquinolone and comparison antibiotics prescriptions per reference population (persons aged ≥18 years) per year

The number of fluoroquinolone prescriptions as well as the comparison antibiotics of the ADR studies (amoxicillin with or without clavulanic acid or cephalosporins) in each year is calculated for each study site and the prescription rate per 100,000 persons is calculated by using the adult (aged 18 or more years) population of that specific year for that specific country and visualized with a line plot. The sources of reference populations are:

- Denmark: Statistics Denmark. Population at the first day of the quarter by sex and age (0 to 125 years), available from 1968 to 2023.
- Finland: Statistics Finland. Population according to age (1-year 0-112) and sex, available from 1972-2022
- Germany: Data from the 2022 census will be available in Summer 2024, extrapolation on statutory health insurance population which is covered by the German claims data source
- Portugal: Statistics Portugal. Resident population (Long series, available from 1970 – 2021) by Sex and Age; Annual – Statistics Portugal, Annual estimates of resident population.

Outcome: proportion of FQ and comparison antibiotics prescriptions out of all oral antibiotic prescriptions

The number of FQ and comparison antibiotics (amoxicillin with or without clavulanic acid or cephalosporins) prescriptions, as well as the number of all oral antibiotic prescriptions per study site are calculated in six-month time windows, and the proportion of FQ prescriptions as well as comparison antibiotics (and 95% CI) of all oral antibiotic prescriptions is calculated and visualized to allow the assessment of between-country differences, temporal change over time and possible effect of regulatory warnings. For FQ analyses, if discontinuity is evident based on these visual inspections, we will perform interrupted time series to quantify the impact of regulatory warnings.

The cutoff (s) for interrupted time series for FQs is decided based on visual inspection of time-trends and the following dates of regulatory warnings are considered for all countries: EMA review initiation date 9.2.2017, EMA warning date 5.10.2018, European Commission dates for legally binding decisions on restriction of use 14.2.2019, 11.3.2019. In addition, the following country-specific days are used if this is necessary based on the visual inspection of the data: Germany: 26.10.2018, 8.4.2019. Finland: review initiation 15.2.2017. Portugal: 4.10.2018, 25.03.2019, 15.10.2020.

Outcome: change in user characteristics

To illustrate the possible change in user characteristics over time, the average age, proportion of women, and prevalence of comorbidities among FQ users and users of comparison antibiotics in each country are calculated in

the same time windows as above (six-months) and visualized by methods developed in WP3. We will illustrate the prevalence of common comorbidities including cardiovascular diseases, stroke, diabetes, asthma/COPD, cancer, dementia as well as urinary infections if the data allows (Table E1).

10.3 Prespecified adverse drug reactions

We will investigate the outcomes in a cohort study of new users of FQ or a comparison antibiotic (amoxicillin with or without clavulanic acid and cephalosporins) and are aged ≥18 years at initiation. The following exclusion criteria are applied to this cohort:

- no purchases of FQ or comparison antibiotics in the preceding 365 days prior to the index date (washout period)
- no hospitalization with infection as the main diagnosis during washout

We will perform separate analyses for the incidence of the following outcomes (prespecified ADRs) using ACNU design:

1. Sudden cardiac death and cardiac arrhythmia
    - Separate sub-study on sudden cardiac death
2. Aortic aneurysm and dissection
3. Acute toxic liver diseases (Drug-induced liver injury)
4. Polyneuropathy (Drug-induced peripheral polyneuropathy)

In each analysis, those with prevalent outcomes (i.e. persons who had a record of the specific outcome before the follow-up as specified in Table D) are excluded from the analyses. The following additional exclusion criteria are used for specific outcomes:

Outcome 1: Sudden cardiac death and cardiac arrhythmia:

Malnutrition, coma, cachexia, dependence on enabling machines and devices, not elsewhere classified (except wheelchair), poisoning by narcotics and psychodysleptics [hallucinogens], sensitivity analyses excluding persons with mental and behavioral disorders due to psychoactive substance use and cancer diagnosis during the previous year. In addition, atrial fibrillation or oral anticoagulant treatment for the composite outcome but not for sudden cardiac death.

Outcome 2: Aortic aneurysm and dissection:

Coma, dependence on enabling machines and devices, not elsewhere classified (except wheelchair), poisoning by narcotics and psychodysleptics [hallucinogens], cancer diagnosis during the previous year.

Outcome 3: Acute toxic liver diseases:

Coma, dependence on enabling machines and devices, not elsewhere classified (except wheelchair), poisoning by narcotics and psychodysleptics [hallucinogens], acute virus hepatitis due to diagnostic problems, obesity. Cancer diagnosis during the previous year. Exclusion of other liver diseases: alcoholic liver disease, chronic hepatic failure and unspecified hepatic failure, chronic hepatitis, fibrosis and cirrhosis of liver, other inflammatory liver diseases, other diseases of liver, liver disorders in diseases classified elsewhere, use of drugs associated with drug-induced liver injury.

Outcome 4: Drug induced peripheral polyneuropathy

Coma, dependence on enabling machines and devices, not elsewhere classified (except wheelchair), poisoning by narcotics and psychodysleptics [hallucinogens], cancer diagnosis during the previous year, obesity, hereditary and idiopathic neuropathy, inflammatory polyneuropathy, other polyneuropathies, alcoholic polyneuropathy, polyneuropathy due to other toxic agents, other specified polyneuropathies unspecified polyneuropathy, and polyneuropathy in diseases classified elsewhere. In addition, sensitivity analyses excluding persons with diabetes are done as neuropathy is a common complication in diabetes.

10.4 Statistical Analyses

In each outcome analysis of the ADR study, the follow-up begins from index date, and ends on day 90, date of death, end of database coverage, or outcome, whichever occurred first. In addition, we will perform sensitivity analyses censoring on the date of exposure crossover if there are such instances in our data. Country-specific age and sex-standardised rates of outcomes per exposure group are calculated using European adult population as the reference population.

Propensity scores are derived from covariates listed in Table E2 using logistic regression and IPT weights are derived with trimming of 2.5% of distribution. Covariate balancing is evaluated by plotting the standardised mean differences of individual covariates before and after the weighting. We will fit IPT-weighted Cox regression (if the assumptions are met) for the entire follow-up, as well as separate models for different 10-day windows. In addition, absolute risks differences and NNH are calculated.

11. Data management and quality control

The OMOP mapping is checked and validated, e.g. with the OHDSI Data Quality Dashboard tool (https://github.com/OHDSI/DataQualityDashboard).
To ensure legal compliance and data privacy preservation, each of the data sources will be accessed in a data privacy preserving manner, and each partner will be responsible for their own data. The Real4Reg will generally follow the paradigm of bringing algorithms to the data rather than the other way around.

Data storage

In Denmark data will be stored on the secure server Forskermaskine of the Danish Health Data Authority. Access is restricted to persons with permission granted by the Danish Health Data Authority and is controlled using a two-step authorisation process.

In Finland the data are stored in the audited remote use environment Kapseli provided by the National Health and Social Data Permit Authority Findata. Access is restricted to persons with permission to use granted by the Findata and controlled by a two-step authorisation process.

In Germany, data will be stored in the Health Data Lab (FDZ) of the Federal Institute for Drugs and Medical Devices (BfArM). The developed algorithms will run on the internal database and only the results will be made available to the researchers.

In Portugal, data will be stored on a server in Infarmed and access will be permitted using an authorisation process.

Independent review

Overall, access to the data is restricted in order to comply with data protection regulations, also see section 12. For external review of analyses, access can be applied for from the data authorities granting the permission (Table A).

Study results will be available for independent review on the Real4Reg website https://www.real4reg.eu/.

12. Limitations

In addition to general limitations of observational designs regarding causal inference, combination of data from different sources may pose limitations. As mentioned in the section 8 Data sources, there is some variability in settings from which the prescriptions are captured, as well as country-specific differences in treatment practices. However, we do not perceive this kind of heterogeneity merely as a limitation, but also an interesting potential contributor to variability in results, and therefore country-specific analyses are conducted instead of pooling the individual-level data.

The IPT-weighting approach can control for measured and recorded confounders. Therefore, residual and/or unmeasured confounding is possible. Application of external adjustment will be considered and conducted if possible. If IPT-weighting does not appropriately control for confounding, we will apply fine stratification weights. Further challenges include coding biases, and nonrandom misclassification. All of these limitations are borne in mind when interpreting and communicating the results, and when evaluating how RWD can be applied in post-authorisation use cases. Although these limitations may affect the robustness of the findings, observations on the aforementioned limitations are also a key outcome of the Real4Reg project and provide important insight on feasibility of using real-world data on regulatory decision-making.

A strength of our data is that we use national healthcare registers from countries with strong public healthcare system, increasing the generalisability of the results.

13. Ethical/data protection issues

Real4Reg is entirely registry based and most of the data sources used in this study are currently already used for pharmaco-epidemiological research. The Real4Reg partners from different EU member states will process personal data from individuals which are collected in national/regional electronic health record databases. Due to the sensitive nature of this personal medical data, we strive to take all reasonable measures to ensure compliance with ethical and regulatory issues on privacy. When required, the study protocols will be reviewed by the national data permit authorities (e.g. Findata) and by Institutional Review Boards (IRBs) of the respective participant institutions and/or data sources.

The pseudonymized patient-level data will not leave any of the data holding organisations. Instead developed algorithms will be brought to the data: Each data holding organisation will set up a dedicated server within in a demilitarised zone, where algorithms can be developed, and calculations can take place. Separate data processing agreements will make sure that neither models nor data can leave these servers, hence providing strong data protection. The intended users of the AI algorithms are statisticians in regulatory agencies or universities. They will be informed about their interaction with AI techniques. Users will be appropriately trained to understand the capabilities, limitations and risks of AI algorithms.

The consortium asserts that all procedures contributing to this research comply with the ethical standards of the relevant national laws of all participating countries and according to the Helsinki Declaration. All consortium partners have a well-developed mechanism to ensure that European and/or local regulations dealing with ethical use of the data and adequate privacy control are adhered to. All data sources will be processed in compliance with relevant legislation and guidance and in line with the General Data Protection Regulation (EU 2016/679). Specifically, Denmark, Finland and Portugal have obtained local approval for their contributions from the respective local Data Protection Agencies. For Germany no direct access to personal data will occur, therefore no approval from a state agency is necessary.

We will statistically attempt to assess whether predictions made by our models are unintentionally impacted by gender. If we find such biases, we will try to eliminate them (e.g. by removing according variables in the training data). If this is unsuccessful, we will raise an according warning.

There are no further ethical risks as models will only be used to support regulatory decision-making, but not replace it. More specifically, our models will provide additional sources of evidence to the regulator, which he/she can consider jointly with the clinical trial data provided by companies.

According to European law, registers and claims data can be used for research without obtaining individual informed consent.

14. Amendments and deviations

This section will document amendments and deviations. Country-specific adjustments of analyses according to data availability and national differences in coding practices may be performed. In Germany, diagnoses are available on a quarterly basis, and a binary variable coding withdrawal from statutory health insurance (which the dataset covers) for reasons of death or change to a private medical insurance is available without specifying the reason for withdrawal. Date or cause of death are not available in Germany.

Updates to the protocol are documented in the appendix (Table A2).

15. Plan for communication of study results

Overall, Real4Reg is committed to a rapid and effective dissemination, exploitation and communication of project results as well as newly generated knowledge to all relevant audiences. This includes the medical, pharmacist and applied regulatory science community as well as all other health care professionals and public health experts including health insurances, regulators, health technology assessment (HTA) and policy makers. Another important target audience will be the general public as the successful implementation and extension of the effective use of RWE in the regulatory and HTA context will require active participation of all patients receiving drug treatments. A dedicated work package WP6 will handle all dissemination tasks.

In brief, the Real4Reg project will:

1. Be promoted online via a public website. This project website will contain information about the overall scope of the project and background, as well as information on individual work packages, use cases, the project team, events, and results. It will also include a section aimed at patients and the general public.

2. Include press releases and a newsletter to raise public awareness of the project as well as social media accounts from the partner organisations (LinkedIn and X (formerly Twitter)).

3. Be regularly presented at international medical and scientific conferences and will be published in well-known peer-reviewed national and international scientific peer-reviewed journals. The Real4Reg consortium embraces the concept of providing open-access whenever possible for a timely dissemination within the scientific and regulatory/ HTA community.

4. Organise events such as workshops and symposia, to publicise the results and their implications for society, including activities dedicated for patients.

16. Timeline

Real4Reg is a four-year project performed during the period January 1, 2023 to December 31, 2026.

Tentative timeline for specific milestones related to WP2 use case 3 are outlined below:

Item	Deliverable/milestone	Month/year
Registration in the EU PAS Register	D	6/2023
Data access & preprocessing finalised for use case 3	M	10/2023
Analyses for changes in broad-spectrum antibiotic use	M	4/2024
Analyses for known ADRs of FQ completed	M	9/2025
Scientific results for use case 3	D	2/2026
Report of good practice examples in postauthorisation RWD use for guidance & training	D	12/2026

17. References

Keskimäki I, Aro S. Accuracy of data on diagnoses, procedures and accidents in the Finnish Hospital Discharge Register. *Int J Health Sciences* 1991;2(1):15–21.

Laugesen K, Ludvigsson JF, Schmidt M, et al. Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries. Clin Epidemiol 2021; 13: 533-54.

Schmidt M, Pedersen L, Sorensen HT. The Danish Civil Registration System as a tool in epidemiology. European Journal of Epidemiology 2014; 29(8): 541-9.

Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 2015; 7: 449-90.

Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. Clin Epidemiol 2019; 11: 563-91.

Sund R. Quality of the Finnish Hospital Discharge Register: A systematic review. Scand J Public Health. 2012 Aug;40(6):505-15.

Sundboll J, Adelborg K, Munch T, Froslev T, Sorensen HT, Botker HE, et al. Positive predictive value of cardiovascular diagnoses in the Danish National Patient Registry: a validation study. BMJ Open. 2016;6(11):e012832.

18. List of Tables

C. Inclusion Criteria

D. Exclusion Criteria for ADR studies: sudden cardiac death, arrhythmia (CDA); aortic aneurysm and dissection (AAS); acute toxic liver disease (ATLD); drug-induced polyneuropathy (DIP), sensitivity outcome sudden cardiac death (SCD)

E. Predefined Covariates

F. Outcome

19. List of appendices

TABLE A1 ANALYSIS SPECIFICATIONS FOR THE ADR STUDIES

TABLE A2 LIST OF PROTOCOL REVISIONS

A. Meta-data about data source and software							
Country	**Data source(s)**	**Description**	**Study period**	**Data extraction date/version**	**Data sampling/ extraction criteria**	**Data linkage**	**Type(s) of data**
Denmark	Danish Civil Registration System (CRS)	Contains individual-level information on personal-identification number (Civil Personal Register – CPR number), demographics, and vital status.	1968-	February 2024	Since 1995	pseudoanonymised CPR-number	Register (administrative, disease register)
	Danish National Prescription Register (NPR)	Contains information on all sales of human and veterinary medicinal products in Denmark.	1995-				
	Danish National Patient Registry (LPR)	Contains information on all hospital contacts.	1977				
	National Hospital Medication Register	Contains information on medication use in public hospitals (currently under development)	2018-				
	Danish Cancer Registry	Contains information on all incident primary cancer diagnoses.	1943, mandatory reporting 1987-				
	Danish Register of Causes of Death	Records on underlying cause of death, and contributory causes	1977				
Portugal	National User Register (RNU)	Contains individual-level information on personal-identifiable number, sex, birth date, residence		Presumably April 2024	Since 2014	RNU (Master Patient Index)	Register
	National Electronic Reimbursed Dispensing Register (CCMSNS)	Contains information on all reimbursed sales of medicines in Portugal					Claims data
	Hospital Morbidity Database (BDMH)	Inpatient and outpatient events and procedures (ICD-10-CM/PCS) in public hospitals				RNU (Master Patient Index)	Register
	Primary care database (BICSP)	Outpatient events and procedures (ICPC-2)				RNU (Master Patient Index)	Register
	Death register database (SICO)	Contains information on death dates, causes				RNU (Master Patient Index)	Register

Germany	Health Data Lab	Contains individual-level information with a personal-identifiable number on all people covered by statutory health insurances (SHI). Information comprises demographics, insurance status and days covered, outpatient medicinal products prescriptions, inpatient and outpatient diagnoses and procedures, further health care sector information (e.g., care status, remedies and aids)	2008-2021	Presumably fall 2024. Possibly preliminary data only Version 1.0	Since 2008	not applicable	Routinely collected health claims
Finland	Care register for health care (Hilmo, AvoHilmo)	Hospitalisations	1972 -	September 2023	2005-2022	by pseudonymised personal identification number	National registers
		Specialised healthcare outpatient visits	1998 -				
		Primary healthcare outpatient visits (diagnoses, procedures, required level of assistance at discharge. Also medications + vaccines in the newer data since ca. 2015 but the completeness of these med/vacc data has not been assessed yet	2011 -				
	Special reimbursements	comorbidity information, reimbursement code & ICD9/10)	1972 -		1972 – 2022 (chronic conditions)		
	Kanta physiological measurements	*measurement type, results, reference values, units…)*	2014 -	September 2023	Since 2014, confirmed results only		
	Kanta laboratory measurements	*measurement type, results, reference values, units…)*	2014 -				
	Dispensed prescriptions, consisting of: Prescription register	ATC, drug name, purchase date, amount, strength, dosing in newer data…	1995 -		2007-2021		
	Kanta electronic prescription database	ATC, drug name, purchase date, amount, strength, dosing in newer data…	2010 -		2010-2023		
	Causes of death register	death dates, causes, how the cause was ascertained	1972 -		2010-2021		

	Statistics Finland socioeconomic information	education, occupation	1972 -		1990-2020		Census and national register

B. Cohort entry defining criterion / Drug utilisation study and prespecified adverse drug reaction (ADR) studies										
Country	**Study population name(s)**	**Day 0 Description**	**Number of entries**	**Washout window**	**Care Setting[1]**	**Code Type**	**Diagnosis position[2]**	**Incident with respect to…**	**Pre-specified**	**Varied for sensitivity**
Denmark	Antibiotic purchaser, drug utilisation study	purchase date from the prescription register	multiple (one per purchase)	none	OP	ATC, 7-characters	n/a	n/a	yes	no
Portugal		Purchase date from the Electronic Prescription and Dispensing Register	multiple (one per purchase)	none	OP	ATC, 7-characters	n/a	n/a	yes	no
Germany		prescription date	multiple (one per purchase)	none		ATC, 7-characters	n/a	n/a	yes	no
Finland		purchase date from prescription register	multiple (one per purchase)	none	OP	ATC, 7-characters	n/a	n/a	yes	no
Denmark, Portugal, Finland	FQ use, ADR study	purchase date from the prescription register/database	multiple (one per purchase)	[-365, -1]	OP	ATC, J01MA	n/a	any J01	yes	no
Denmark, Portugal, Finland	comparator use, ADR study	purchase date from the prescription register/database	multiple (one per purchase)	[-365, -1]	OP	ATC, J01DB, J01DC, J01DD, J01DE, J01CA04, J01CR02	n/a	any J01	yes	no
Germany	FQ use, ADR study	prescription date	multiple (one per purchase)	[-365, -1]		ATC, J01MA	n/a	any J01	yes	no
Germany	comparator use, ADR study	prescription date	multiple (one per purchase)	[-365, -1]	OP	ATC, J01DB, J01DC, J01DD, J01DE, J01CA04, J01CR02	n/a	any J01	yes	no

[1] Please enter all that apply. Valid entries: IP = inpatient, OP = outpatient, ED = emergency department, any, other, n/a = not applicable.

[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

C. Inclusion Criteria							
Criterion	**Details**	**Order of application**	**Assessment window**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**	**Source for algorithm**
Observable		1	[-365; 0] [0; 90] [0; 365]	ADR studies	yes	yes (0; 90 main analysis and 0; 365 sensitivity analysis)	Specifically developed
Age ≥18 years on index date	prescription date of J01 purchase in drug utilisation study, date on purchase/prescription of FQ/comparator in the ADR study	2	Day 0	Drug utilisation, ADR studies	yes	no	Specifically developed

D. Exclusion Criteria for ADR studies: sudden cardiac death, arrhythmia (CDA); aortic aneurysm and dissection (AAS); acute toxic liver disease (ATLD); drug-induced polyneuropathy (DIP), sensitivity outcome sudden cardiac death (SCD)								
Criterion	**Order of application**	**Assessment window**	**Care Settings[1]**	**Code Type**	**Diagnosis position[2]**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**
Prevalent antibiotic use	1	[-365; -1]	OP	ATC	n/a	CDA, AAS, SCD, ATLD, DIP	yes	Washout length?
Prevalent outcome	2	[-365; -1]	OP, IP	ICD10	any	CDA	yes	Period length?
Prevalent outcome	2	[-365; -1]	OP, IP	ICD10	any	SCD	yes	Period length?
Prevalent outcome	2	[-365; -1]	OP, IP	ICD10	any	AAS	yes	Period length?
Prevalent outcome	2	[-365; -1]	OP, IP	ICD10	any	ATLD	yes	Period length?
Prevalent outcome	2	[-365; -1]	OP, IP	ICD10	any	DIP	yes	Period length?
Coma	3	[-365; -1]	IP	ICD10	any	CDA, SCD, AAS, ATLD, DIP	yes	Period length?
Dependence on enabling machines and devices	4	[-365; -1]	OP, IP	ICD10	any	CDA, SCD, AAS, ATLD, DIP	yes	Period length?
Poisoning by narcotics and psychodysleptics [hallucinogens]	5	[-365; -1]	OP, IP	ICD10	any	CDA, SCD, AAS, ATLD, DIP	yes	Period length?
Acute virus hepatitis	6	[-365; -1]	OP, IP	ICD10	any	ATLD	yes	Period length?
Malnutrition, cachexia	6	[-365; -1]	OP, IP	ICD10	any	CDA	yes	Period length?
Hereditary and idiopathic neuropathy, inflammatory polyneuropathy, other polyneuropathies, alcoholic polyneuropathy, Polyneuropathy due to other toxic agents, Other specified polyneuropathies Polyneuropathy, unspecified and Polyneuropathy in diseases classified elsewhere	6	[-365; -1]	OP, IP	ICD10	any	DIP	yes	Period length?
Drugs with known hepatotoxicity: specific antiepileptics, analgetics, immune modulators, immune checkpoint inhibitors...	7	[-90; -1]	OP	ATC	any	ATLD	yes	Period length-365 (assess whether this changes the n of excluded)
Chronic liver diseases		[-365; -1]	OP	ICD10	any	ATLD		
Mental and behavioural Disorders due to psychoactive substance use	7	[-365; -1]	OP, IP	ICD10	any	CDA, SCD	yes	Additional exclusion for sensitivity analyses

Cancer	8	[-365; -1]	IP	ICD10	any	CDA, AAS, SCD, ATLD, DIP	yes	Additional exclusion for sensitivity analyses
Oral anticoagulant treatment	9	[-365; -1]	OP	ATC	n/a	CDA	yes	
Atrial fibrillation	10	[-365; -1]	OP	ICD10	any	CDA	yes	
Diabetes	7	[-365; -1]	OP	ATC	n/a	DIP	yes	Additional exclusion for sensitivity analyses
Diabetes	7	[-365; -1]	OP, IP	ICD10	any	DIP	yes	Additional exclusion for sensitivity analyses
Obesity	8	[-365; -1]	OP, IP	ICD10	any	DIP, ATLD		
Hospitalisation during washout		[-365; -1]	IP	ICD10	Main	all		Sensitivity analyses without this exclusion

E1. Predefined Covariates in drug utilization study							
Characteristic	**Details**	**Type of variable**	**Assessment window**	**Care Settings[1]**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**
Sex		Binary	0	n/a	Drug utilization study	Yes	No
Age		Continuous	0	n/a	Drug utilization study	Yes	No
Cardiovascular diseases		Binary	Since available data until antibiotic prescription	OP/IP	Drug utilization study	Yes	No
Stroke		Binary	Since available data until antibiotic prescription	OP/IP	Drug utilization study	Yes	No
Diabetes		Binary	Since available data until antibiotic prescription	OP/IP	Drug utilization study	Yes	No
Asthma/COPD		Binary	Since available data until antibiotic prescription	OP/IP	Drug utilization study	Yes	No
Dementia		Binary	Since available data until antibiotic prescription	OP/IP	Drug utilization study	Yes	No
Cancer		Binary	Since available data until antibiotic prescription	OP/IP	Drug utilization study	Yes	No
Urinary infections		Binary	-365; antibiotic prescription	OP	Drug utilization study	Yes	No

E2. Predefined Covariates in the ADR studies							
Characteristic	**Details**	**Type of variable**	**Assessment window**	**Care Settings[1]**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**
Sex		Binary	0	n/a	all	Yes	No
Age		Continuous	0	n/a	all	Yes	No
Socioeconomic position	Based on highest occupational social class until index date, used in those countries where it is available from	Ordinal	0	n/a	all	Yes	No
Cardiovascular disease medications		Binary	[-1825; -1]	OP	all	Yes	No
Ischemic heart disease		Binary	[-1825; -1]	OP, IP	all	Yes	No
Hypertension		Binary	[-1825; -1]	OP, IP	all	Yes	No
Heart failure/cardiomyopathy		Binary	[-1825; -1]	OP, IP	all	Yes	No
Valve disorders		Binary	[-1825; -1]	OP, IP	all	Yes	No
Cerebrovascular disease		Binary	[-1825; -1]	OP, IP	all	Yes	No
Arterial disease		Binary	[-1825; -1]	OP, IP	all	Yes	No
Cardiac procedures and surgeries during washout		Binary	[-356; -1]	IP	all	Yes	No
Lung disease		Binary	[-1825; -1]	OP, IP	all	Yes	No
Liver disease		Binary	[-1825; -1]	OP, IP	all	Yes	No
Liver procedures (biopsy, transplantation)		Binary	[-1825; -1]	OP, IP	all	Yes	No
Renal disease		Binary	[-1825; -1]	OP, IP	all	Yes	No
Rheumatic disease		Binary	[-1825; -1]	OP, IP	all	Yes	No
Schizophrenia and Mood disorders		Binary	[-1825; -1]	OP, IP	all	Yes	No
Dementia		Binary	[-1825; -1]	OP, IP	all	Yes	No
Diabetes		Binary	[-1825; -1]	OP, IP	all	Yes	No
Antipsychotics		Binary	[-1825; -1]	OP	all	Yes	No
Antidepressants		Binary	[-1825; -1]	OP	all	Yes	No
Anxiolytic, hypnotic or sedative		Binary	[-1825; -1]	OP	all	Yes	No
Oral glucocorticoid		Binary	[-1825; -1]	OP	all	Yes	No

Non-study antibiotic use or any antibiotic use in year preceding the washout			[-731; -366]	OP	all	Yes	No
Hospital days during washout		Continuous	[-356; -1]	IP	all	Yes	No
Number of outpatient visits to specialized		Continuous	[-356; -1]	OP	all	Yes	No
Number of outpatient visits to primary healthcare during washout		Continuous	[-356; -1]	OP	all	Yes	No
malnutrition or cachexia, arrythmia		Binary	[-1825; -1]	OP, IP	DIP, ATLD	Yes	No
obesity		Binary	[-1825; -1]	OP, IP	all except DIP, ATLD	Yes	No
oesophageal varices		Binary	[-1825; -1]	OP, IP	ATLD	Yes	No
liver disease		Binary	[-1825; -1]	OP, IP	ATLD	Yes	No

F. Outcome Outcome name	Outcome measurement characteristics	Type of outcome	Washout window	Care Settings[1]	Code Category	Diagnosis position[2]	Applied to study populations:	Pre-specified	Varied for sensitivity
annual prescription	number of any oral antibiotic prescriptions /reference population per year	Continuous	none	OP	-	n/a	Drug utilisation	yes	no
annual prescription	number of any oral fluoroquinolone prescriptions /reference population per year	Continuous	none	OP		n/a	Drug utilisation	yes	no
annual prescription	number of any oral amoxicillin or amoxicillin with clavulanic acid prescriptions /reference population per year	Continuous	none	OP		n/a	Drug utilisation	yes	no
annual prescription	number of any oral cephalosporin prescriptions /reference population per year	Continuous	none	OP		n/a	Drug utilisation	yes	no
proportion	Proportion of fluoroquinolone prescriptions / all oral antibiotic prescriptions per six-months	Continuous [0,1]	none	OP	-	n/a	Drug utilisation	yes	no
proportion	Proportion of amoxicillin or amoxicillin with clavulanic acid prescriptions / all oral antibiotic prescriptions per six-months	Continuous [0,1]	none	OP		n/a	Drug utilisation	yes	no
proportion	Proportion of cephalosporin prescriptions / all oral antibiotic prescriptions per six-months	Continuous [0,1]	none	OP		n/a	Drug utilisation	yes	no
Sudden cardiac death, arrhythmias		binary		IP, OP	ICD10, hospitalisations and causes of death	any	CDA	yes	no
Sudden cardiac death		binary		IP, OP	ICD10, hospitalisations and causes of death	any	SCD	yes	no
Aortic aneurysms and dissections		binary		OP, IP	ICD10, hospitalisations	any	AAS	yes	no

					and causes of death				
Acute toxic liver disease		binary		IP	ICD10, hospitalisations and causes of death	any	ATLD	yes	no
Polyneuropathy		binary		IP	ICD10, hospitalisations	any	DIP	yes	no

TABLE A1 ANALYSIS SPECIFICATIONS FOR THE ADR STUDIES		
	Primary	**Secondary 1**
Hypothesis:	n/a	n/a
Study population(s)	FQ or comparison antibiotic initiators	FQ or comparison antibiotic initiators
Outcome:	prespecified ADRs (table F), 90-day ITT	prespecified ADRs (table F), 365-day ITT
Software:	R/Python/SAS	R/Python/SAS
Model(s):	crude, adjusted	crude, adjusted
Confounding adjustment method		
Bivariate		
Multivariable	IPTW	IPTW
Other		
(specify details)		
Missing data methods	Not applicable	Not applicable
Missing indicators		
Complete case		
Last value carried forward		
Multiple imputation (specify variables)		
Other (please specify)		
Subgroup Analysis	site-specific analyses conducted	site-specific analyses conducted

TABLE A2 LIST OF PROTOCOL REVISIONS

Section	Change (marked red)	Comment	Date of revision
All sections	Correcting grammar mistakes	Grammar mistakes	28.3.2024
3	Furthermore, where data permits in each partner, the change in age and sex ~~sociodemographic characteristics~~ distribution and prevalence of comorbidities in FQ users, and amoxicillin or cephalosporin users over time will be assessed.	Only age and sex	12.3.2024
3	(Figure 1). Figure 1. Visualisation of study design in the ADR studies.	Adding a reference to the figure below the text and specifying the Figure text.	12.3.2024
4&5	Inclusion and exclusion criteria are described in Tables C~~B~~ and D~~C~~. ….in Tables B and C~~E~~ and…	Wrong Table names	12.3.2024
5	Exposure definition and measurement	Making a bolded header into header with number → Header 5. Following header numbers updated accordingly to Headers and to the text.	
5	Alternative approaches such as time-dependent exposure modelling are considering if the number of repeated prescriptions is significant.	Adding a clarification	12.3.2024
6	The proportion of FQ and comparison antibiotics (amoxicillin with or without clavulanic acid or cephalosporins) prescriptions of all oral antibiotic prescriptions ~~per year~~ The ~~trend~~change in user characteristics over the study period	Adding a clarification	12.3.2024
7	In the ADR studies, confounding…	Adding a clarification	12.3.2024
9.1	Outcomes are identified from hospital encounter registers in Finland and Denmark and additionally from primary care data in Finland ~~and primary care visits in Finland and Denmark~~. In Portugal, disease and treatment registers (including primary care) ~~in Portugal~~ and claims in Germany are used. In addition, causes of death register is used in all other countries except Germany. Covariates are identified from the dispensing (ATC codes) codes from the prescription databases and diagnosis (ICD-10, ICPC-2) and procedure (NOMESCO, ICD-10-CM/PCS) codes	Adding a clarification	26.3.2024

	from the healthcare registers and databases described above.		
10.2	**Outcome: fluoroquinolone and comparison antibiotics prescriptions per** reference population (persons aged ≥18 years) per year~~100,000 persons~~ The number of fluoroquinolone prescriptions as well as the comparison antibiotics of the ADR studies (amoxicillin with or without clavulanic acid or cephalosporins) in each year is calculated… Germany: Data from the 2022 census will be available in Summer 2024 ~~Microcensus available from 2011 or 2022 (depending on availability),~~, extrapolation on statutory health insurance population which is covered by the German claims data source **Outcome: proportion of FQ and comparison antibiotics prescriptions out of all oral antibiotic prescriptions** The number of FQ and comparison antibiotics (amoxicillin with or without clavulanic acid or cephalosporins) prescriptions,…. …and the proportion of FQ prescriptions as well as comparison antibiotics (and 95% CI) of…. For FQ analyses, if discontinuity is evident…	Adding a clarification	12.3.2024
10.2	To illustrate the possible change in user characteristics over time, the average age, proportion of women, and prevalence of comorbidities among FQ users and users of comparison antibiotics ~~on the purchase date~~ in each country are calculated in the same time windows as above (six-months) and visualized by methods developed in WP3. We will illustrate the prevalence of common comorbidities including ~~such as~~ cardiovascular diseases, stroke, diabetes, asthma/COPD, cancer, dementia as well as urinary infections if the data allows (Table E1).	Adding a clarification	12.3.2024
10.3	Obesity	Adding obesity to acute toxic liver diseases and drug induced peripheral polyneuropathy according to Table D	12.3.2024
10.4	…as well as separate models for different 10-day windows. In addition, absolute risks differences and NNH are calculated.	Adding a clarification	26.3.2024

12	If IPT-weighting does not appropriately control for confounding, we will apply fine stratification weights.	Adding a clarification	12.3.2024
13	We will statistically attempt to assess whether predictions made by our models are unintentionally impacted by ~~ethnicity and~~ gender.	Deleted ethnicity as none of the countries have that information in the data	26.3.2024
14	Updates to the protocol are documented in the appendix (Table A2).	Clarification	26.3.2024
16	Changes in timeline Analyses for changes in broad-spectrum antibiotic use from 2/2024 into 4/2024 Analyses for known ADRs of FQ completed from 5/2024 into 9/2025	Changes in the data access	12.3.2024
Table A	Updates in data extraction date/version and Data sampling/ extraction criteria Finland: Care register for health care (Hilmo, AvoHilmo)	Updates	12.3.2024
Tables	Any J01~~A~~	Change into proper ATC-code; J01A was for tetracyclines	12.3.2024
Table B	Code type: ATC, J01DB, J01DC, J01DD, J01DE, J01CA04, J01CR02 Finland: purchase date from ~~Kanta/~~prescription register	Amoxicillin(-clavunate) was missing Update to use of only prescription register data in Finland for drug utilization study	28.3.2024
Table C	~~Drug utilization,~~ ADR studies	Clarification	12.3.2024
Table D	~~Chardiac arrythmia~~ (CDA)	Change for oral anticoagulant treatment and for atrial fibrillation according to exclusion criteria in the protocol text	12.3.2024
Table E1&E2	Previous Table E into Table E2 as new Table E1 was added	Adding a new table for covariates in drug utilization studies and change in Table header accordingly	26.3.2024
Table E2	Based on highest occupational social class~~/ income on~~ until index date, used in those countries where it is available from		
Table F	Adding other intended outcome analyses for drug utilization study number of any oral antibiotic prescriptions /reference population per year Proportion of fluoroquinolone prescriptions / all oral antibiotic prescriptions per ~~year~~ six-months	Specifying analyses	12.3.2024
Tables	Deleting text in grey rows (instruction for filling the tables)	Deleting unnecessary information	12.3.2024

Table A1	TABLE A1 ANALYSIS SPECIFICATIONS FOR THE ADR STUDIES Deleting the column “secondary 2”	Focus in the table is on the ADR studies, not drug utilization study	26.3.2024
Table A2	TABLE A2 LIST OF PROTOCOL REVISIONS	Adding a new table that lists all relevant protocol updates	26.3.2024

Title	Work Package (WP) 2, Use Case 4 **Using Real World Data to evaluate effectiveness of SGLT-2 inhibitors in heart failure (drug repurposing)**
Protocol version identifier	1.1
Date of last version of protocol	02-04-2024
EU PAS register number	EUPAS105544
Active substance	Sodium-Glucose Cotransporter-2 (SGLT-2) inhibitors (ATC A10BK)
Medicinal product	Not specified
Product reference	Not specified
Procedure number	Not specified
Marketing authorisation holder)	Not specified
Joint PASS	No
Research question and objectives	The overall objective is the preparation of a good practice example for safety analyses of Real-World Data (RWD) for the post-authorisation stage. We assess how RWD from four European countries can be used to generate high-quality, population-based information on the effectiveness evaluation in drug repurposing by using Sodium-Glucose Cotransporter-2 (SGLT-2) inhibitors and heart failure-related outcomes as the use case. The specific objectives are to: 1. Describe national time trends in the prescription of individual new oral antidiabetic drugs, such as SGLT-2 and dipeptidyl peptidase-4 (DPP-4) inhibitors, from 2012 to 2022, applying the longest available period depending on data availability in each participating country. 2. Evaluate the comparative effectiveness of SGLT-2 inhibitors by emulating a clinical trial on the effect of SGLT-2 inhibitors on heart failure-related and all-cause outcomes (hospitalisations and mortality) with an active comparator new user design (ACNU) using DPP-4 inhibitors as the active comparator. 3. Describe similarities and differences between available data sources from the participating countries and evaluate whether heterogeneity in data leads to heterogeneity in results and how this should be taken into account in reporting. 4. To assess how RWD can be used to generate high-quality,

	population-based information on benefits and to evaluate additional value of Artificial Intelligence (AI) in clinical trial emulation
Countries of study	Denmark, Finland, Germany, Portugal
Author Members of Work Packages 1-3 of the Real4Reg project and associates contributed to the protocol and adopt it:	Anna-Maija Tolppanen, Sirpa Hartikainen, Anne Paakinaho Aborageh, Mohamed Adewuyi, Davis Arzideh, Roxana Becker, Cornelia Bräuner, Elvira Braithwaite, Billy Costa, Inês Ehrenstein, Vera Fernandes, Joana Froehlich, Holger Furtado, Cláudia Haenisch, Britta Hartikainen, Sirpa Heß, Steffen Horváth-Puhó, Erzsébet Kallio, Aleksi Kjær, Jesper Korcinska Handest, Monika Roberta Nagy, Dávid Paakinaho, Anne Peltner, Jonas Pfeifer, Kerstin Pylkkanen, Liisa Roethlein, Christoph Russek, Martin Rajamaki, Blair Schneider, Katharina Silva, Célia Tolppanen, Anna-Maija Vancraeyenest, Aurélie Vo, Huu Thuan Wicherski, Julia

Table of Contents for WP2, Use Case 4

1. Abbreviations

ACNU Active Comparator New User

AI Artificial Intelligence

ATC Anatomical Therapeutic Chemical

CDM common data model

CI confidence interval

COPD chronic obstructive pulmonary disease

DPP-4 dipeptidyl peptidase-4EMA European Medicines Agency

HTA Health Technology Assessment

ICD international classification of diseases

IRB Institutional Review Board

IPTW inverse probability of treatment weighting

ML Machine Learning

NNT number needed to treat

OMOP Observational Medical Outcomes Partnership

SGLT-2 Sodium-Glucose Cotransporter-2

QC Quality Control

RWD Real-World Data

WP work package

2. Research question

The overall objective of WP2 is the preparation of good practice examples for drug safety (Use Case 3) and effectiveness (Use Case 4) analyses of Real-World Data (RWD) for the drug post-authorisation stage. This protocol describes the analyses for Use Case 4 that evaluates the effectiveness of Sodium-Glucose Cotransporter-2 (SGLT-2) inhibitors on heart failure-related outcomes, which are an important public health concern. The target population of this study are adults who receive a prescription of SGLT-2 inhibitor.

2.1 Study objectives

To evaluate how RWD can be used to generate high-quality, population-based information for post-authorisation effectiveness studies by using SGLT-2 inhibitors as an example.

The specific aims of Use Case 4 are to:

1. Describe national time trends in the use of individual new oral antidiabetics SGLT-2 and dipeptidyl peptidase-4 (DPP-4) inhibitors in the period from 2012 to 2022, applying the longest available period depending on data availability in each participating country. SGLT-2 inhibitors were introduced to the market in late 2012 in Denmark and Finland, in 2013 in Germany, and in 2014 in Portugal. The first DPP-4 inhibitors (sita- and vildagliptin) were approved by the European Medicines Agency (EMA) in 2007.
2. Evaluate the comparative effectiveness of SGLT-2 inhibitors by emulating a clinical trial on the effect of SGLT-2 inhibitors on heart failure-related and all-cause hospitalisations and mortality with active comparator new user design (ACNU) using DPP-4 inhibitors as the comparator.
3. Describe similarities and differences between available data sources from four participating countries to examine whether heterogeneity in data leads to heterogeneity in results and how this should be considered in reporting.
4. To evaluate how RWD can be used to generate high-quality, population-based information on benefits, and to evaluate additional value of Artificial Intelligence (AI) in trial emulation.

2.2 Specific tasks

1. Data preprocessing
    - 1.1 Provide data access and carry out data pre-processing tasks (Quality Control [QC] and conversion to Observational Medical Outcomes Partnership [OMOP]-common data model [CDM])
    - 1.2 Provide a meta data catalogue and summary.

2. Perform a descriptive drug utilisation study of national trends in SGLT-2 inhibitor and DPP-4 inhibitor utilisation in relation to overall use of non-insulin antidiabetic drugs in 2012-2022, applying the longest available period depending on data availability in each participating country
3. Perform outcome analyses for heart failure-related and all-cause hospitalisations and mortality by target trial emulation
    3.1 comparison of common vs site-specific propensity scores
4. Assessment of data heterogeneity and whether this leads to heterogeneity of results

3. Study design

The study uses secondary data. Data sources from the participating countries (Denmark, Finland, Germany and Portugal) are listed in Table A. Two study designs will be applied:

A descriptive drug utilization study that aims to illustrate the changes in prescription retrievals of SGLT-2 inhibitors and DPP-4 inhibitors in relation to non-insulin antidiabetic drugs during the study period (2012-2022), applying the longest available period depending on data availability in each participating country. We will assess the proportion of users of these drugs per all non-insulin antidiabetic drug users in 6-month intervals of the study period. Furthermore, where data permits, the change in the distribution of age, sex, and prevalence of comorbidities in SGLT-2 and DPP-4 inhibitor users over time will be ascertained to evaluate whether the reports on cardiovascular benefits of SGLT-2 inhibitors led to change in user characteristics over time.

A cohort study with multiple eligibility-based entries per person using an ACNU design to assess all-cause and heart failure- related hospitalisations and death among SGLT-2 inhibitor users compared with DPP-4 inhibitors users (comparator) (Figure 1). The incidence of heart failure among the exposure groups will be calculated and the relative risk between the groups will be estimated using hazard ratios. The accumulation of hospitalisations and hospital days per person-year between the groups is compared. Finally, absolute risk difference and number needed to treat (NNT) for heart failure diagnosis will be calculated.

4. Source and study populations

The source populations are adult populations with diabetes (≥ 40 years on the date of index prescription retrieval) of Denmark, Finland, Germany, and Portugal. The study participants are persons with written (Germany) or purchased prescription (Denmark, Finland, Germany, Portugal) for non-insulin antidiabetic

drugs in 2012-2021, identified from national healthcare registers (Denmark, Finland and Portugal) and claims data (Germany) (Tables B1 and B2).

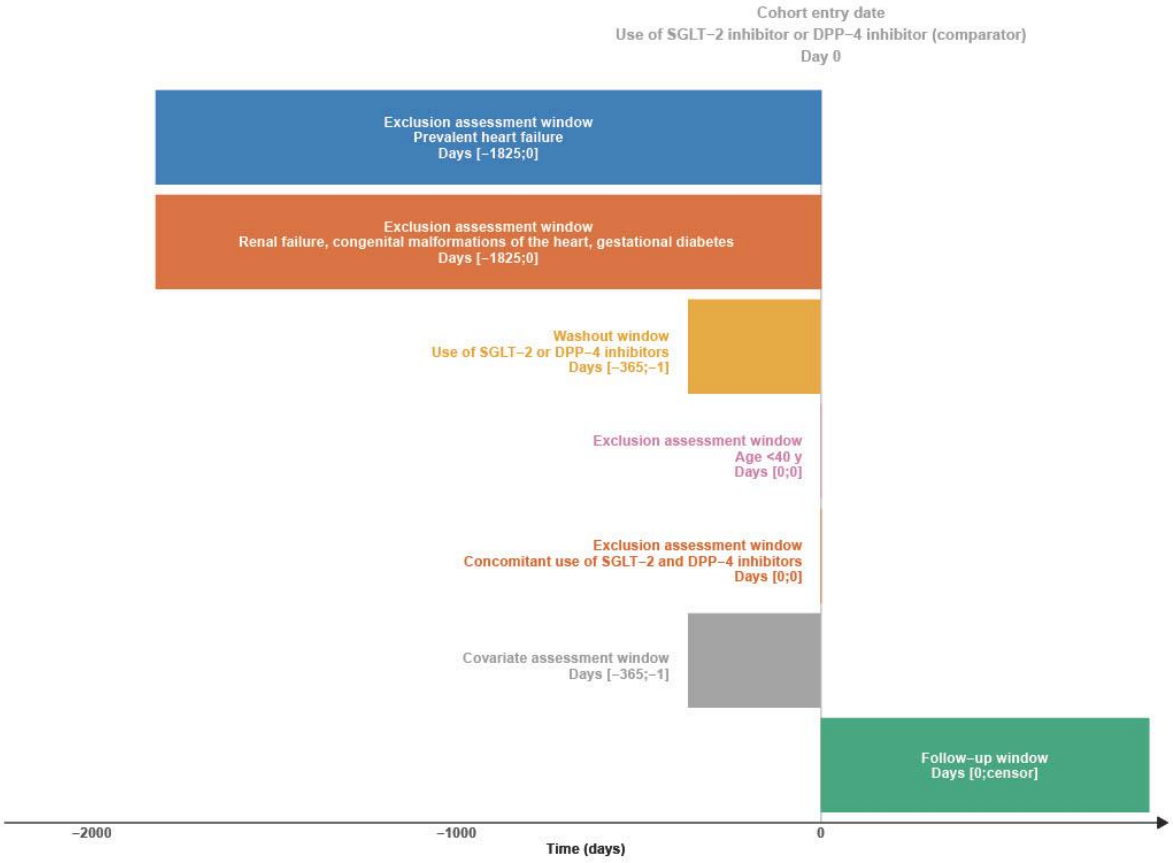

Figure 1. Visualization of the study design for the effectiveness studies.

5. Exposure definition and measurement

In the effectiveness studies, the exposure is defined as new use of SGLT-2 inhibitor or the comparator (DPP-4 inhibitor) (excluding saxagliptin and alogliptin due to the FDA warning on their associations with increased risk of heart failure) using a one-year washout as described in Tables B2 and C (Figure 1). Concomitant users of SGLT-2 and DPP-4 inhibitors are excluded. Included persons are permitted to have multiple entries provided that each exposure initiation meets the inclusion criteria. Drug exposure periods are modelled using fixed assumptions or waiting time distribution modelling. The modelling assumptions are decided after initial data checks. If there is exposure crossover (a SGLT-2 inhibitor user initiates use with DPP-4 inhibitor or vice versa), or initiation of saxa- or alogliptin, the follow-up ends on that day. Initiations

of other medications (antidiabetic or cardiovascular) during the follow-up are allowed. We will also add a carryover period to the end of exposure assessment, the length to be decided later. Based on earlier studies the benefits seem to remain relatively long, even six months after discontinuation. The follow-up begins on index date and ends on death, end of data linkage period, exposure change or migration away from database, whichever occurs first.

6. Outcome definition and measurement

Outcomes are listed in Table F. Drug utilisation study outcomes (the proportion of exposure and comparator drug users out of all non-insulin antidiabetic drug users) are calculated separately for all four countries and identified from the prescription databases of each country. Dispensed prescriptions are used for Denmark, Finland and Portugal, written prescriptions for Germany (see Chapter 9.1). The following outcomes are calculated for the drug utilisation study:

- The proportion of SGLT-2 and DPP-4 inhibitor users of all non-insulin antidiabetic users per 6-months
- The trends in DPP-4 and SGLT-2 inhibitor user characteristics over time

The heart failure-specific and all-cause outcomes (hospitalisation and mortality) are identified from the hospital discharges and outpatient visits in all countries. In addition, causes of death are used in Denmark, Finland, and Portugal. Incident outcomes are identified using a five-year washout period. The following outcomes are assessed using the same study population:

- hospitalisation, any cause (binary)
- number of hospital days per person-year, any cause
- number of hospital days per person-year, with heart failure as discharge diagnosis
- incident diagnosis of heart failure
- death with heart failure recorded in death certificate
- all-cause mortality

Exclusion criteria are listed in detail in Table D. In addition to prevalent users of SGLT-2 or DPP-4 inhibitors, persons with renal failure, congenital malformations of the heart, gestational diabetes or those with previous heart failure are excluded.

The validity of the Danish and Finnish healthcare registers has been reviewed previously (Laugesen et al 2021, Schmidt et al 2014, 2015, 2019, Sund 2012), and a high positive predictive value for diseases of the circulatory system diagnoses (Keskimäki and Aro 1991) and heart failure (Mähönen et al 2013, Vuori et al 2020) in Finland as well as cardiovascular diagnoses in Denmark (Sundboll et al 2016) has been reported.

7. Bias

Confounding by indication will be controlled by the utilisation of the ACNU design. We have selected another antidiabetic drug comparator that is typically used as an add-on therapy and has the same administration route as SGLT-2 inhibitors. In addition, we will use inverse probability of treatment weighting (IPTW) to control for confounding. The propensity scores from which the weights are derived, are calculated based on predefined covariates (Table E). In addition, a Machine Learning (ML) approach is used to derive the propensity score and to explore, in how far the causal effect of the drug exposure can be predicted on the basis of historical patient-level data prior to index date.

Prevalent user bias is controlled by using a one-year washout to identify new use. We expect the misclassification of exposure to be relatively small and nondifferential between the exposure groups. It is possible that heart failures are under recorded in the data sources we use, but the use of outpatient diagnosis data from primary and specialized healthcare enables us to detect higher proportions than mere restriction to inpatient data.

Misclassification of outcome may also occur due to differences in coding practices, and it is possible that there is between-country and within-country variation in recording heart failure. However, we assume this will be nondifferential between the exposure groups, and therefore it is not expected to impact relative risk. However, absolute risk estimate may be affected. Between-country variation in recording heart failure will be considered as an explanation if there is significant between-country variation in the incidence rates. We expect that there will be limited data available on the severity of heart failure or diabetes. The latter may result in residual confounding, although we will use different proxies for diabetes severity (used antidiabetic medication, duration of diabetes, hospitalisations with diabetes recoded as a diagnosis code). In addition, we will use information on laboratory measures such as HbA1c if these are available/recorded in sufficient amount (available from Finland and Denmark only).

The three countries that use causes of death register data have very similar processes for ascertaining the cause of death and registration of them. The European Commission Regulation (EC) No 1338/2008 confirms the variables, specifications and metadata which the EU Member States have to supply concerning the statistics on causes of death. In Finland, Denmark and Portugal, death certificates are issued by the physician who establishes the death and causes are recorded using the International Classification of Diseases (ICD) -10 coding. In Finland, most causes of death are based on the clinical data, and autopsied are

confirmed in less than 20% of cases (14.6% deaths confirmed with forensic and 3.6% by medical autopsy in 2020). If the information in the death certificate is deficient, inconsistent or difficult to classify, a clarification request is issued.

8. Effect measure modification

We will evaluate whether the associations between SGLT-2 inhibitor and outcomes are similar across sites by describing differences between the incidence rates, absolute and relative risk estimates and their 95% Confidence Intervals (CI). No other effect modification analyses are planned unless we observe a time trend in covariates in the drug utilisation study.

9. Data sources

Data sources are described in Table A.

9.1 Data sources and coding system for exposure, outcomes and covariates

Denmark, Finland, and Portugal have excellent data linkage resources at the national level. In Germany, only one data source is used. All countries will provide RWD from national healthcare registers or claims data. Data sources used to ascertain information on exposure, outcomes and covariates are listed in Tables B, E, F.

Exposure to non-insulin antidiabetic drugs is identified from dispensed prescriptions (all for Denmark, reimbursed for Finland and Portugal) and written prescriptions according to claims data (Germany) using Anatomical Therapeutic Chemical (ATC) codes. The dates of dispensing (Denmark, Finland, Portugal) or prescription writing (Germany) days are used to ascertain the beginning of follow-up. Germany has data on prescribed drugs available until 2018 and on both prescribed and dispensed drugs from 2019 onwards; data on prescribed drugs only are used for Germany to be consistent.

There is variation for settings from which the drug use is captured. The Danish data cover prescriptions purchased from community-dwelling settings and drugs administered in hospitals. The Finnish and Portuguese data sources cover prescriptions purchased in community-dwelling settings and residential care, but not drugs administered in hospitals. The German data cover claims from community-dwelling and long-term care settings.

Outcomes are identified from hospital encounter registers in Finland and Denmark, and additionally from primary care data in Finland. In Portugal, disease and treatment registers (including primary care) and

claims in Germany are used. In addition, causes of death register is used in all other countries except Germany.

Covariates are identified from the dispensing (ATC codes) codes from the prescription databases and diagnosis (ICD-10, International Classification of Primary Care, 2nd edition (ICPC-2)) and procedure (NOMESCO, ICD-10-CM/PCS) codes from the healthcare registers and databases described above.

9.2 Linkage method between data

In Denmark and Finland, the data are linked based on pseudonymised unique person identifiers. In Portugal, the linkage is performed using a Master Patient Index (through the User Number, available for all citizens) followed by a pseudonymised unique person identifier. No linkage will occur in Germany as only one data source is applied.

10. Analysis plan

Country-specific adjustments of analyses according to data availability and national differences in coding practices may be performed. We estimate that over 3 million people used SGLT-2 inhibitors or DPP-4 inhibitors in the four countries during the study period.

10.1 Data preprocessing

Data are inspected for missing values, coding errors and erratic dates and converted to the OMOP CDM.

10.2 Drug utilisation study

Outcome: proportion of SGLT-2 inhibitors users and DPP-4 inhibitors users out of all non-insulin antidiabetic drug users

The number of SGLT-2 inhibitor users, as well as the number of all non-insulin antidiabetic drug users per study site are calculated in six-month time windows, and the proportion of SGLT-2 inhibitor users (and 95% CI) of all non-insulin antidiabetic users is calculated and visualized to allow the assessment of between-country differences and change over time.

Outcome: change in user characteristics

To illustrate the possible change in user characteristics over time, the average age, proportion of women, and prevalence of comorbidities among SGLT-2 and DPP-4 inhibitor users on the purchase date in Denmark, Finland, and Portugal and prescription date in Germany are calculated in the same time windows as above and visualized by methods developed in WP3. We will illustrate the prevalence of common comorbidities

(ischemic heart disease, heart failure, hypertension, stroke, and renal insufficiency), use of other diabetes medication categories in the preceding year, and duration of diabetes (in countries with available data).

10.3 Outcomes for effectiveness analyses

In each country, we will evaluate the effectiveness outcomes in one cohort using ACNU design. Separate analyses for each outcome are performed. Specifications of outcomes are listed in Table F.

10.4 Statistical analysis for effectiveness studies

The follow-up begins from index date (day 0), with multiple entries allowed per person provided that each entry meets the inclusion and exclusion criteria. The exposure is defined as incident use of SGLT-2 inhibitor or the comparator (DPP-4) inhibitor (excluding saxagliptin and alogliptin) using a one-year washout. Concomitant users of SGLT-2 and DPP-4 inhibitors during washout, or on day 0 are excluded. Drug exposure periods are modelled using fixed assumptions or waiting time distribution modelling. The modelling assumptions are decided after initial data checks. The follow-up ends on outcome, death, migration out of the country and if there is exposure crossover (SGLT-2 inhibitor initiates with DPP-4 inhibitor or vice versa), initiation of saxa- or alogliptin, or end of study period, whatever comes first. Initiations of other medications (antidiabetic or cardiovascular) during the follow-up are allowed. We will also add a carryover period to the end of exposure assessment, the length is decided later based on expert opinion.

Propensity scores are derived from covariates which will be chosen based on up-to date literature review and listed to Table E using logistic regression and IPT weights are derived with trimming of 2.5% of distribution. Covariate balancing is evaluated by plotting the standardised mean differences of individual covariates before and after the weighting.

We will fit IPT-weighted Cox regression (if the assumptions are met) for the entire follow-up for incident heart failure and mortality outcomes. In addition, absolute risks differences are calculated to calculate NNT.

Hospitalisation rates are calculated as sum of hospital days or admissions per person-year and rates between SGLT-2 and DPP-4 inhibitor initiators are compared with negative binomial regression (if appropriate for distribution) using IPT weights.

11. Data management and quality control

To ensure legal compliance and data privacy preservation, each of the data sources will be accessed in a data privacy preserving manner, and each partner will be responsible for their own data. The Real4Reg will generally follow the paradigm of bringing algorithms to the data rather than the other way around.

Data storage

In Denmark data will be stored on the secure server Forvaltningsmaskine of the Danish Health Data Authority. Access is restricted to persons with permission granted by the Danish Health Data Authority and is controlled using a two-step authorisation process.

In Finland the data are stored in the audited remote use environment Kapseli provided by the National Health and Social Data Permit Authority Findata. Access is restricted to persons with permission to use granted by the Findata and controlled by a two-step authorisation process.

In Germany, data will be stored in the Health Data Lab (FDZ) of the Federal Institute for Drugs and Medical Devices (BfArM). The developed algorithms will run on the internal data base and only the results will be made available to the researchers.

In Portugal, data will be stored on a server in Infarmed and access will be permitted using an authorisation process.

Independent review

Overall, access to the data is restricted to comply with data protection regulations, also see section 13. For external review of analyses, access can be applied for from the data authorities granting the permission (Table A), but access cannot be guaranteed.

Study results will be available for independent review on the Real4Reg website: www.real4reg.eu.

Quality control

We expect the amount of missing data to be minor as the completeness of the data sources and specific field we apply in this study question is close to 100% based on our experience. Meta-data will be mapped to the OMOP common data model. The OMOP mapping is checked and validated, e.g. with the OHDSI Data Quality Dashboard tool (https://github.com/OHDSI/DataQualityDashboard).

12. Limitations

In addition to general limitations of observational designs regarding causal inference, combination of data from different sources may pose limitations. As mentioned in the section 9 Data sources, there is some variability in settings from which the prescriptions are captured, as well as country-specific differences in treatment practices. However, we do not perceive this kind of heterogeneity merely as a limitation, but also an interesting potential contributor to variability in results, and therefore country-specific analyses are conducted instead of pooling the individual-level data.

The IPT-weighting approach can control for measured and recorded confounders. Therefore, residual and/or unmeasured confounding is possible. Application of external adjustment will be considered and conducted if possible. If IPT-weighting does not appropriately control for confounding, we will apply fine stratification weights. Further challenges include coding biases, and nonrandom misclassification. All of these limitations are considered when interpreting and communicating the results, and when evaluating how RWD can be applied in post-authorisation use cases. Although these limitations may affect the robustness of the findings, observations on the aforementioned limitations are also a key outcome of the Real4Reg project and provide important insight on feasibility of using real-world data on regulatory decision-making.

A strength of our data is that we use national healthcare registers from countries with strong public healthcare system, increasing the generalisability of the results.

13. Ethical/data protection issues

Real4Reg is entirely registry based and most of the data sources used in this study are currently already used for pharmaco-epidemiological research. The Real4Reg partners from different EU member states will process personal data from individuals which are collected in national/regional electronic health record databases. Due to the sensitive nature of this personal medical data, we strive to take all reasonable measures to ensure compliance with ethical and regulatory issues on privacy. When required, the study protocols will be reviewed by the national data permit authorities (e.g. Findata) and by Institutional Review Boards (IRBs) of the respective participant institutions and/or data sources.

The pseudonymized patient-level data will not leave any of the data holding organisations. Instead developed algorithms will be brought to the data. Each data holding organisation will set up a dedicated server within a demilitarised zone, where algorithms can be developed, and calculations can take place. Separate data processing agreements will make sure that neither models nor data can leave these servers, hence providing strong data protection. The intended users of the AI algorithms are statisticians in regulatory agencies or universities. They will be informed about their interaction with AI techniques. Users will be appropriately trained to understand the capabilities, limitations, and risks of AI algorithms.

The consortium asserts that all procedures contributing to this research comply with the ethical standards of the relevant national laws of all participating countries and according to the Helsinki Declaration. All consortium partners have a well-developed mechanism to ensure that European and/or local regulations dealing with ethical use of the data and adequate privacy control are adhered to. All data sources will be processed in compliance with relevant legislation and guidance and in line with the General Data Protection Regulation (EU 2016/679). Specifically, Denmark, Finland and Portugal have obtained local approval for their contributions from the respective local Data Protection Agencies. For Germany no direct access to personal data will occur, therefore no approval from a state agency is necessary.

We will statistically attempt to assess whether predictions made by our models are unintentionally impacted by gender. If we find such biases, we will try to eliminate them (e.g. by removing according variables in the training data). If this is unsuccessful, we will raise an according warning.

There are no further ethical risks as models will only be used to support regulatory decision-making, but not replace it. More specifically, our models will provide additional sources of evidence to the regulator, which he/she can consider jointly with the clinical trial data provided by companies.

According to European law, registers and claims data can be used for research without obtaining individual informed consent.

14. Amendments and deviations

This section will document amendments and deviations. Country-specific adjustments of analyses according to data availability and national differences in coding practices may be performed. In Germany, diagnoses are available on a quarterly basis, and a binary variable coding withdrawal from statutory health insurance

(which the dataset covers) for reasons of death or change to a private medical insurance is available without specifying the reason for withdrawal. Date or cause of death are not available in Germany.

Updates to the protocol are documented in the appendix (Table A2).

15. Plan for communication of study results

Overall, Real4Reg is committed to a rapid and effective dissemination, exploitation and communication of project results as well as newly generated knowledge to all relevant audiences. This includes the medical, pharmacist and applied regulatory science community as well as all other health care professionals and public health experts including health insurances, regulators, health technology assessment (HTA) and policy makers. Another important target audience will be the general public as the successful implementation and extension of the effective use of RWE in the regulatory and HTA context will require active participation of all patients receiving drug treatments. A dedicated work package, WP6, will handle all dissemination tasks.

In brief, the Real4Reg project will:

1. Be promoted online via a public website. This project website contains information about the overall scope of the project and background, as well as information on individual work packages, the project team events, and results. It also includes a section aimed at patients and the general public.

2. Include press releases and a newsletter to raise public awareness of the project as well as social media accounts from the partner organisations (LinkedIn and X (formerly Twitter)).

3. Be regularly presented at international medical and scientific conferences and will be published in well-known peer-reviewed national and international scientific peer-reviewed journals. The Real4Reg consortium embraces the concept of providing open-access whenever possible for a timely dissemination within the scientific and regulatory/ HTA community.

4. Organise events such as workshops and symposia, to publicise the results and their implications for society, including activities dedicated for patients.

16. Timeline

The project is a four-year project performed during the period January 1, 2023 to December 31, 2026.

Tentative timelines for specific milestones related to WP2 use case 4 are outlined below:

Item	Deliverable/milestone	Month/Year
Registration in the EU PAS Register	D	6/2023
Data access & preprocessing finalised for use case 4	M	10/2023
Analyses for changes in non-insulin antidiabetic drug use completed	M	9/2025
Analyses of comparative effectiveness of SGLT-2 inhibitors completed	M	2/2026
Scientific results for use case 4	D	12/2026
Report of good practice examples in postauthorisation RWD use for guidance & training	D	12/2026

17. References

Keskimäki I, Aro S. Accuracy of data on diagnoses, procedures and accidents in the Finnish Hospital Discharge Register. *Int J Health Sciences* 1991;2(1):15–21.

Laugesen K, Ludvigsson JF, Schmidt M, et al. Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries. Clin Epidemiol 2021; 13: 533-54.

Mähönen M, Jula A, Harald K, Antikainen R, Tuomilehto J, Zeller T, et al. The validity of heart failure diagnoses obtained from administrative registers. *Eur J Prev Cardiol* 2013 Apr;20(2):254-9.

Schmidt M, Pedersen L, Sorensen HT. The Danish Civil Registration System as a tool in epidemiology. European Journal of Epidemiology 2014; 29(8): 541-9.

Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 2015; 7: 449-90.

Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. Clin Epidemiol 2019; 11: 563-91.

Sund R. Quality of the Finnish Hospital Discharge Register: A systematic review. Scand J Public Health. 2012 Aug;40(6):505-15.

Sundboll J, Adelborg K, Munch T, Froslev T, Sorensen HT, Botker HE, et al. Positive predictive value of cardiovascular diagnoses in the Danish National Patient Registry: a validation study. BMJ Open. 2016;6(11):e012832.

Vuori MA, Laukkanen JA, Pietilä A, Havulinna AS, Kähönen M, Salomaa V, Niiranen TJ; FinnGen investigators. The validity of heart failure diagnoses in the Finnish Hospital Discharge Register. Scand J Public Health. 2020 Feb;48(1):20-28.

18. List of Tables

19. Appendices

Table A. Meta-data about data source and software

A. Meta-data about data source and software									
Country	**Data source(s)**	**Description**	**Study period**	**Data extraction date/version**	**Data sampling/ extraction criteria**	**Data linkage**	**Type(s) of data**	**Data conversion**	**Software to create study population**
Denmark	Danish Civil Registration System (CRS)	Contains individual-level information on personal-identification number (Civil Personal Register – CPR number), demographics, and vital status.	1968-	February 2024	Since 1995	CPR-number (pseudoanonymised)	Registry (administrative, disease registry)	OMOP	R/SAS
	Danish National Prescription Register (NPR)	Contains information on all sales of human and veterinary medicinal products in Denmark.	1995-						
	Danish National Patient Registry (LPR)	Contains information on all hospital contacts.	1977						
	National Hospital Medication Register	Contains information on medication use in public hospitals (currently under development)	2018-						
	Danish Cancer Registry	Contains information on all incident primary cancer diagnoses.	1943, mandatory reporting 1987-						

	Danish Register of Causes of Death	Records on underlying cause of death, and contributory causes	1977						
Portugal	National User Register (RNU)	Contains individual-level information on personal-identifiable number, sex, birth date, residence		April 2024	Since 2014	RNU (Master Patient Index)	Registry	OMOP	R/Python
	National Electronic Reimbursed Dispensing Register (CCMSNS)	Contains information on all reimbursed sales of medicines in Portugal				RNU (Master Patient Index)	Claims data		
	Hospital Morbidity Database (BDMH)	Inpatient and outpatient events and procedures (ICD-10-CM/PCS) in public hospitals				RNU (Master Patient Index)	Registry		
	Primary care databse (BICSP)	Outpatient events and procedures (ICPC-2)				RNU (Master Patient Index)	Registry		
	Death register database (SICO)	Contains information on death dates, causes				RNU (Master Patient Index)	Registry		

Germany	Health Data Lab	Contains individual-level information with a personal-identifiable number on all people covered by statutory health insurances (SHI). Information comprises demographics, insurance status and days covered, outpatient medicinal products prescriptions, inpatient and outpatient diagnoses and procedures, further health care sector information (e.g., care status, remedies and aids)	2008-2021	Presumable Fall 2024 Possibly preliminary data only version 1.0	Since 2008	not applicable	Routinely collected health claims	OMOP	R/Python
Finland	Care register for health care	Hospitalisations	1972 -	September 2023	2010-2022	by pseudonymised personal identification number	National registers	OMOP	R / SAS
		Specialised healthcare outpatient visits	1998 -						
		General healthcare outpatient visits (diagnoses, procedures, required level of assistance at discharge. Also medications + vaccines in the newer data since ca. 2015 but the completeness of these med/vacc data has not been assessed yet	2011 -						
	Special reimbursements	comorbidity information, reimbursement code & ICD9/10)	1972 -		1972 - 2022(chronic conditions)				

	Kanta physiological measurements	*measurement type, results, reference values, units…)*	2014 -	September 2023	Since2014, confirmed results only				
	Kanta laboratory results	*measurement type, results, reference values, units…)*	2014 -						
	Dispensed prescriptions, consisting of: Prescription register	ATC, drug name, purchase date, amount, strength, dosing in newer data…	1995 -	September 2023	2010-2021				
	Kanta electronic prescription database	ATC, drug name, purchase date, amount, strength, dosing in newer data…	2010 -		2010-2023				
	Causes of death register	death dates, causes, how the cause was ascertained	1972 -		2014-2021				
	Statistics Finland socioeconomic information	education, occupation	1972 -		1990-2020		Census and national register		

Table B1 Cohort entry defining criterion / Drug utilisation study

B1. Cohort entry defining criterion / Drug utilisation study												
Country	**Study population name(s)**	**Day 0 Description**	**Number of entries**	**Type of entry**	**Washout window**	**Care Setting[1]**	**Code Type**	**Diagnosis position[2]**	**Incident with respect to…**	**Pre-specified**	**Varied for sensitivity**	**Source of algorithm**
Denmark	Noninsulinic antidiabetic drug purchaser	purchase date from the Danish prescription register	multiple (one per purchase)		none	OP, IP	ATC, 7-characters	n/a	n/a	yes	no	
Portugal	Noninsulinic antidiabetic drug purchaser	Purchase date from the Portuguese Electronic Prescription and Dispensing Register	multiple (one per purchase)		none	OP	ATC, 7-characters	n/a	n/a	yes	no	
Germany	Noninsulinic antidiabetic drug purchaser	prescription date	multiple (one per purchase)		none		ATC, 7-characters	n/a	n/a	yes	no	
Finland	Noninsulinic antidiabetic drug purchaser	purchase date from Kanta/prescription register	multiple (one per purchase)		none	OP	ATC, 7-characters	n/a	n/a	yes	no	

[1] Please enter all that apply. Valid entries: IP = inpatient, OP = outpatient, ED = emergency department, any, other, n/a = not applicable. See Appendix E for details on how care setting is defined

[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter

Table B2 Cohort entry defining criterion / Effectiveness studies

B2. Cohort entry defining criterion / Effectiveness studies												
Country	**Study population name(s)**	**Day 0 Description**	**Number of entries**	**Type of entry**	**Washout window**	**Care Setting[1]**	**Code Type**	**Diagnosis position[2]**	**Incident with respect to…**	**Pre-specified**	**Varied for sensitivity**	**Source of algorithm**
Denmark, Portugal, Finland	SGLT-2 inhibitor use	purchase date from the prescription register/database	multiple (one per purchase)		[-365, -1]	OP	ATC	n/a	SGLT-2 inhibitor and DPP-4 inhibitor	yes	no	
Denmark, Portugal, Finland	comparator use (DPP-4)	purchase date from the prescription register/database	multiple (one per purchase)		[-365, -1]	OP	ATC	n/a	SGLT-2 inhibitor and DPP-4 inhibitor	yes	no	
Germany	SGLT-2 inhibitor use	prescription date	multiple (one per purchase)		[-365, -1]		ATC	n/a	SGLT-2 inhibitor and DPP-4 inhibitor	yes	no	
Germany	comparator use (DPP-4)	prescription date	multiple (one per purchase)		[-365, -1]	OP	ATC	n/a	SGLT-2 inhibitor and DPP-4 inhibitor	yes	no	

[1] Please enter all that apply. Valid entries: IP = inpatient, OP = outpatient, ED = emergency department, any, other, n/a = not applicable. See Appendix E for details on how care setting is defined

[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

Table C Inclusion criteria

C. Inclusion Criteria										
Criterion	**Details**	**Order of application**	**Assessment window**	**Care Settings[1]**	**Code Type**	**Diagnosis position[2]**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**	**Source for algorithm**
Observable		1	[-365, 0]	n/a	n/a	n/a	effectiveness studies	yes	no	Specifically developed
Age ≥40 years on index date		2	Day 0	n/a	n/a	n/a	effectiveness studies	yes	no	Specifically developed

[1] Please enter all that apply. Valid entries: IP = inpatient, OP = outpatient, ED = emergency department, any, other, n/a = not applicable. See Appendix E for details on how care setting is defined

[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

Table D Exclusion Criteria for Effectiveness studies

D. Exclusion Criteria for Effectiveness studies										
Criterion	**Details**	**Order of application**	**Assessment window**	**Care Settings[1]**	**Code Type**	**Diagnosis position[2]**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**	**Source for algorithm**
Exposed to SGLT-2 or comparator during washout	ATC	1	[-365; -1]	OP	ATC	n/a	Effectiveness	yes	Washout length?	
Exposed to SGLT2-DPP-4 inhibitor combination during washout	ATC	2	[-365; -1]	OP	ATC	any	Effectiveness	yes	Period length?	
Exposed to saxagliptin / alogliptin during washout	ATC	3	[-365; -1]	OP	ATC	any	Effectiveness	yes	Period length?	
Prevalent heart failure	ICD-10	4	-1825, -1]	IP	ICD-10	Any	Effectiveness	yes	Period length?	
Renal insufficiency	ICD-10	5	-1095, -1]	IP, OP	ICD-10	Any	Effectiveness	yes	Period length?	
Congenital malformations of heart	ICD-10	6	-1825, -1]	IP, OP	ICD-10	Any	Effectiveness	yes	Period length?	

[1] Please enter all that apply. Valid entries: IP = inpatient, OP = outpatient, ED = emergency department, any, other, n/a = not applicable. See Appendix E for details on how care setting is defined

[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

Table E Predefined Covariates

E. Predefined Covariates										
Characteristic	**Details**	**Type of variable**	**Assessment window**	**Care Settings[1]**	**Code Type**	**Diagnosis position[2]**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**	**Source for algorithm**
Sex		Binary	0	n/a		n/a	all	Yes	No	
Age		Continuous	0	n/a		n/a	all	Yes	No	
Socioeconomic position	highest occupational social class until index date, used in those countries where it is available from	ordinary		n/a			all	Yes	No	

TO BE DEFINED LATER

Table F Outcome

f. Outcome											
Outcome name	**Outcome measurement characteristics**	**Primary outcome ?**	**Type of outcome**	**Washout window**	**Care Settings[1]**	**Code Category**	**Diagnosis position[2]**	**Applied to study populations:**	**Pre-specified**	**Varied for sensitivity**	**Source of algorithm**
proportion	Proportion of SGLT-2 inhibitor prescriptions / all non-insulin diabetes drug prescriptions per year		Derived	n/a	OP		n/a	Drug utilisation	yes	no	
hospital admission, all cause		no	binary	n/a	IP		n/a	effectiveness	yes	no	
number of hospital days, per person-year, any cause		no	continuous	n/a	IP		n/a		yes	no	
number of hospital days, per person-year, with heart failure as discharge diagnosis	ICD-10 I50, I110, I130, I132, J81	yes	continuous	[-1825, -1]	IP		any		yes	no	

Diagnosis of for heart failure	ICD-10 I50, I110, I130, I132, J81	yes	binary, incident	[-1825, -1]	IP/OP		any		yes	no	
Death with heart failure recorded in death certificate	ICD-10 I50, I110, I130, I132, J81	yes	binary, incident	[-1825, -1]	n/a		any		yes	no	
Heart failure diagnosis or death due to heart failure	ICD-10 I50, I110, I130, I132, J81	yes	binary, incident	[-1825, -1]	IP/OP		any		yes	no	
all-cause mortality		no	binary	n/a	n/a		n/a		yes	no	

Abbreviations for study populations: sudden cardiac death, arrhythmia (CDA); aortic aneurysm and dissection (AAS); acute toxic liver disease (ATLD); drug-induced polyneuropathy (DIP), sensitivity outcome sudden cardiac death (SCD)

[1] Please enter all that apply. Valid entries: IP = inpatient, OP = outpatient, ED = emergency department, any, other, n/a = not applicable. See Appendix E for details on how care setting is defined

[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

APPENDIX

Table A1 Analysis specifications

TABLE A1 ANALYSIS SPECIFICATIONS for the effectiveness studies		
	Primary	**Secondary 1**
Hypothesis:	to compare the risk of heart-failure-elated hospitalisations and mortality between SGLT-2 and DPP-4 inhibitor users	to compare the risk of all-cause hospitalisations and mortality between SGLT-2 and DPP-4 inhibitor users
Study population(s)	new users of SGLT-2 of DPP-4 inhibitor	new users of SGLT-2 of DPP-4 inhibitor
Outcome:	hospitalisation for heart failure, hospital days due to heart failure per person-year, death with heart failure recorded as a cause	hospitalisations, accumulation of hospital days, death
Software:	R/Python/SAS	R/Python/SAS
Model(s):	crude, adjusted	crude, adjusted
Confounding adjustment method		
Bivariate		
Multivariable	IPTW	IPTW
Other		
(specify details)		
Missing data methods	Not applicable	Not applicable
Missing indicators		
Complete case		
Last value carried forward		
Multiple imputation (specify variables)		
Other (please specify)		
Subgroup Analysis	site-specific analyses conducted	site-specific analyses conducted

Table A2 LIST OF PROTOCOL REVISIONS

Section	Change (marked red)	Comment	Date of revision
All sections	Correcting grammar mistakes and spelling harmonisation	changed for spelling for continuity; other spelling and grammar mistakes corrected	2024-03-12
1, 2.1	approved by the European Medicines Agency (EMA) in 2007	wrote out the acronym and added to the abbreviation list	2024-03-12
3	We will assess the proportion of prescriptions for these drugs per all non-insulin antidiabetic drugs in 6-month intervals ~~each year~~ of the study period	time intervals updated	2024-03-26
3	Furthermore, where data permits, the change in the distribution of ~~sociodemographic characteristics~~ age, sex and prevalence of comorbidities in SGLT-2 and DPP-4 inhibitor users…	Only age and sex	2024-03-26
3	The incidence of heart failure among the exposure groups will be calculated and the relative risk ~~difference~~ between the groups will be estimated using hazard ratios	changed for clarity	
5	Tables B~~1~~2 and C	wrong table referenced	2024-03-12
	Drug utilisation study outcomes (the proportion of exposure and comparator drug ~~prescriptions~~ users out of all non-insulin antidiabetic drug ~~prescriptions~~ users) are calculated separately for all four countries and identified from the prescription databases of each country.	changed for clarity	2024-03-26
6	The proportion of SGLT-2 and DPP-4 inhibitor ~~prescriptions~~ users of all non-insulin antidiabetic ~~prescriptions~~ users per 6-months~~year~~	Changed for clarity; wrong time frame	2024-03-12
6, 17	(Mähönen et al 2013, Vuori et al 2020); Vuori MA, Laukkanen JA, Pietilä A, Havulinna AS, Kähönen M, Salomaa V, Niiranen TJ; FinnGen investigators. The validity of heart failure diagnoses in the Finnish Hospital Discharge Register. Scand J Public Health. 2020 Feb;48(1):20-28.	reference added	2024-03-12
6	~~However, it should be noted that the previous validation studies in Finland were performed for inpatient registers and the introduction of specialised healthcare polyclinic visits in 1998 and primary healthcare use since 2011 have likely led to improved validity for conditions that can be treated and diagnosed in outpatient settings.~~	Vuori et al 2020 was based on outpatient data	2024-03-12
8	We will evaluate whether the associations between	Changed for	2024-03-26

	SGLT-2 inhibitor and outcomes are similar across sites by describing differences between ~~comparing~~ the incidence rates,…	clarity	
9.1	… outcomes and covariates are listed in Tables B, ~~C, D,~~ E, F.	Wrong tables	2024-03-26
9.1	Outcomes are identified from hospital encounter registers ~~and primary care visits~~ in Finland and Denmark, and additionally from primary care data in Finland. In Portugal, disease and treatment registers (including primary care)~~in Portugal~~ and claims in Germany are used. In addition, causes of death register is used in all other countries except Germany. Covariates are identified from the dispensing (ATC codes) codes from the prescription databases and diagnosis (ICD-10, International Classification of Primary Care, 2nd edition (ICPC-2)…	Changed for clarity	2024-03-26
10	Country-specific adjustments of analyses according to data availability and national differences in coding practices may be performed. We estimate that over 3 million people used SGLT-2 inhibitors or DPP-4 inhibitors in the four countries during the study period.	added for clarity	2024-03-26
10.2	We will illustrate the prevalence of common comorbidities ~~such as~~ (ischemic heart disease, heart failure, hypertension, stroke, and renal insufficiency), use of other diabetes medication categories in the preceding year, and duration of diabetes (in countries with available data). ~~cardiovascular diseases, asthma/ Chronic obstructive pulmonary disease (COPD), cancer, dementia.~~	Comorbidities were agreed upon	2024-03-12
10.2	**Outcome: proportion of SGLT-2 inhibitors users ~~prescriptions~~ and DPP-4 inhibitors users out of all non-insulin antidiabetic users ~~prescriptions~~**	Changed for clarity	2024-03-26
10.3	In each country, we will evaluate the effectiveness outcomes in one cohort using ACNU design.	updated for clarity	2024-03-26
10.3	~~In addition to prevalent users of SGLT-2 or DPP-4 inhibitors, persons with renal failure, congenital malformations of the heart, gestational diabetes, or those with previous heart failure are excluded.~~	repeatitve text, written in section 6	2024-03-26
11	~~Data are inspected for missing values, coding errors and erratic dates and converted to the OMOP CDM.~~	Removed for repetitiveness in same paragraph	2024-03-26
12	If IPT-weighting does not appropriately control for confounding, we will apply fine stratification weights.	added for clarity	2024-03-26
13	We will statistically attempt to assess whether predictions made by our models are unintentionally impacted by ~~ethnicity and~~ gender.	removed due to data availability	2024-03-26
14	Updates to the protocol are documented in the appendix (Table A2)	added for clarity	2024-03-12
15	(LinkedIn and X (formerly Twitter))	updated the name of	2024-03-12

		social media platform	
16	Tentative timelines for specific milestones related to WP2 use case 4 are outlined below:	updated for clarity	2024-03-26
18	D. Exclusion Criteria for Effectiveness studies~~:~~ ~~exposure to SGLT-2 or comparator during washout, exposure to SGLT2-DPP-4 inhibitor combination during washout, Exposure to saxagliptin / alogliptin during washout, prevalent heart failure, renal insufficiency, congenital malformations of heart~~	removed for clarity	2024-03-12
18	~~F~~G. Outcome	Wrong table label	2024-03-12
Table A	Data extraction date/ version and Data sampling/ extraction criteria fields updated	updated for all countries	2024-03-12
Table B2	Incident with respect to… ~~any J01A~~ SGLT-2 inhibitor and DPP-4 inhibitor	wrong use case	2024-03-12
Table C	Assessment window column~~: [-365, 90; 0]~~ [-365,0] Applied to study column**:** ~~Drug utilisation~~, effectiveness studies	Updated for clarity	2024-03-26
Table A1	**TABLE A1 ANALYSIS SPECIFICATIONS for the effectiveness studies** **Column Secondary 2 deleted**	added for clarity; column removed because it was referring to drug utilisation studies	2024-03-26