Janssen Research & Development*

Study Protocol for Retrospective Observational Studies Using Secondary Data

Comparative Risk Assessment of Severe Uterine Bleeding Following Exposure to Direct Oral Anticoagulants: A Network Study Across 4 US Observational Databases

Protocol PCSCVM002566

* Janssen Research & Development (Janssen R&D) is a global organization that operates through different legal entities in various countries. Therefore, the legal entity acting as the sponsor for studies of Janssen R&D may vary. The term "sponsor" is used throughout the protocol to represent these various legal entities.

Status: Approved
Date: 13 April 2020
Prepared by: Janssen Research & Development, LLC
EDMS number: EDMS-RIM-30477, 2.0

TABLE OF CONTENTS

1. LIST OF ABBREVIATIONS

Abbreviation	Description of Abbreviated Term
CAD	Coronary artery disease
CDM	Common data model
CI	Confidence interval
DOAC	Direct oral anticoagulant
DVT	Deep vein thrombosis
EHR	Electronic health record
FDA	Food and Drug Administration
FXa	Factor Xa
HR	Hazard ratio
ITT	Intent-to-treat
NVAF	Non-valvular atrial fibrillation
OMOP	Observational Medical Outcomes Partnership
PAD	Peripheral artery disease
PE	Pulmonary embolism
PPV	Positive predictive value
PS	Propensity score
RR	Relative risk
SMD	Standardized mean difference
SUB	Severe uterine bleeding
THR	Total hip replacement
TKR	Total knee replacement
US	United States
VTE	Venous thromboembolism
Abbreviation	Description of Abbreviated Term
FOIA	Freedom of Information Act

2. RESPONSIBLE PARTIES

2.1. Investigator(s) and Authors

James Weaver – Janssen Global Epidemiology
Zhong Yuan – Janssen Global Epidemiology
Lu Wang – Janssen Global Epidemiology
Azza Shoaibi – Janssen Global Epidemiology
Gowtham Rao – Janssen Global Epidemiology
Leila Larbi – Janssen Global Medical Safety
Amy Freedman- Janssen Global Medical Safety
Elliot Barnathan – Janssen Research & Development
Joanne Foody – Janssen Research & Development
Shujian Wu – Janssen Research & Development
Purve Patel – Janssen Research & Development
Peter Wildgoose – Janssen Pharmaceuticals, Medical Affairs

2.2. Sponsor

Global Epidemiology, Janssen Research & Development, LLC.
1125 Trenton-Harbourton Rd, Titusville NJ 08560

3. ABSTRACT

Rivaroxaban and other anticoagulants are associated with increased bleeding risk because of their biological mechanisms. One clinical trial reported an increased risk of abnormal uterine bleeding for rivaroxaban relative to enoxaparin/warfarin. Findings for abnormal uterine bleeding have been inconsistent among clinical trials for other direct oral anticoagulants. Real-world evidence on the risk of severe uterine bleeding among direct head-to-head comparison of direct oral anticoagulants and warfarin is limited. This study will evaluate severe uterine bleeding for rivaroxaban, apixaban, dabigatran, and warfarin therapies among patients with non-valvular atrial fibrillation, venous thromboembolism, and total hip replacement or total knee replacement surgeries in routine clinical practice. Empirical evaluation has assessed the study design *a priori* for its ability to generate valid comparative risk estimates and informs which results will be reported.

4. AMENDMENTS AND UPDATES

Number	Date	Section of Study Protocol	Amendment or Update	Reason
1	2020/02/10	7.4. Outcome(s) of Interest	Combined the 2 outcome definitions into single outcome	Insufficient event counts for severe uterine bleed with transfusion management
2	2020/02/10	6.2. Secondary Objective(s); 7.3. Study Population(s)	Removed the subgroup analysis sections	Insufficient exposure subpopulation sample sizes and outcome counts
3	2020/02/28	7.4. Outcome(s) of Interest	Removed pruritic rash, pruritus of skin, ataxia as sequela of cerebrovascular disease, and palpitations from negative control outcomes	Potentially causally related to DOACs
4	2020/02/28	6.1. Primary Objective(s); 7.3. Study Population(s)	Removed CAD/PAD indication group from objectives	Apixaban and dabigatran not approved for this indication and other comparisons failed diagnostics

5. RATIONALE AND BACKGROUND

Rivaroxaban is an oral, direct factor Xa (FXa) inhibitor anticoagulant that inhibits thrombus generation and thrombus formation. Rivaroxaban has demonstrated a favorable benefit-risk profile in many clinical settings and is approved in adults in the United States (US) and worldwide for the prevention and/or treatment of multiple thrombosis-mediated conditions, including:

- to reduce the risk of stroke and systemic embolism in patients with nonvalvular atrial fibrillation
- for the treatment of deep vein thrombosis (DVT)

- for the treatment of pulmonary embolism (PE)
- for the reduction in the risk of recurrence of DVT and/or PE in patients at continued risk for recurrent DVT and/or PE after completion of initial treatment lasting at least 6 months
- for the prophylaxis of DVT, which may lead to PE in patients undergoing knee or hip replacement surgery
- for the prophylaxis of venous thromboembolism (VTE) in acutely ill medical patients at risk for thromboembolic complications not at high risk of bleeding
- in combination with aspirin, to reduce the risk of major cardiovascular events (cardiovascular (CV) death, myocardial infarction (MI) and stroke) in patients with chronic coronary artery disease (CAD) or peripheral artery disease (PAD)

Due to the biological mechanism of action, as with other anticoagulants, rivaroxaban is associated with an increased risk of bleeding [http://www.janssenlabels.com/package-insert/product-monograph/prescribing-information/XARELTO-pi.pdf]. In the ROCKET AF trial that randomized more than 14,000 patients with atrial fibrillation (rivaroxaban, N=7,111; warfarin, N=7,125), while the overall ISTH major bleeding rates (%/year) were similar between the treatment groups (rivaroxaban 3.6 vs. warfarin 3.5; hazard ratio [HR]=1.04 (95% confidence interval [CI]: 0.90, 1.20)), gastrointestinal bleeding rate (%/year) was higher in the rivaroxaban group as compared with the warfarin group (rivaroxaban 2.0 vs. warfarin 1.2, HR; HR= 1.61 (95% CI: 1.30, 1.99)).

Abnormal uterine bleeding (AUB) events were also observed in clinical studies. Based on EINSTEIN DVT and PE clinical trials, Martinelli and colleagues[1] evaluated 1737 women younger than 60 and reported that abnormal uterine bleeding occurred more often in rivaroxaban-treated women compared with enoxaparin/VKA-treated women (HR=2.1; 95% CI: 1.6, 2.9). However, findings of AUB from clinical studies have been inconsistent for other direct oral anticoagulants (DOACs, e.g., apixaban and edoxaban) or direct thrombin inhibitors (dabigatran). In the AMPLIFY trial[2], clinically relevant non-major (CRNM) vaginal bleeding occurred in 28 (2.5 %, [28/1122]) apixaban and 24 (2.1 %, [24/1106]) enoxaparin/warfarin recipients (OR=1.2, 95% CI: 0.7–2.0), although CRNM was more likely to be of vaginal origin with apixaban (45%) than with warfarin (20%). Based on the Hokusai-VTE study of women younger than 50 years of age (N=1,233), Scheres and colleagues[3] reported that the rate of abnormal vaginal bleeding was 15/100 person-years (95% CI: 11–19) in women receiving edoxaban and 9/100 person-years (95% CI: 6–12) in the warfarin group (hazard ratio: 1.7, 95% CI: 1.1–2.5). Based on the pooled analysis of the RE-COVER and RE-MEDY trials[4] (n = 1280), in which 643 received direct thrombin inhibitor dabigatran and 637 received warfarin, dabigatran showed an lower AUB rate of 5.9%, compared to 9.6% with VKA (OR 0.6; 95% CI 0.4-0.9).

Despite findings from clinical trials, data from routine clinical practice are generally lacking. The current voluntary, sponsor initiated study is intended to evaluate severe uterine bleeding, defined as vaginal bleed plus transfusion or surgical management, in relation to rivaroxaban, apixaban, dabigatran, and warfarin therapies. To aggregate treatment exposure in similar clinical settings, the above indications are grouped into the following categories: non-valvular atrial fibrillation (NVAF), Venous thromboembolism (VTE, including deep vein thrombosis (DVT) and pulmonary

embolism (PE)), and total hip replacement or total knee replacement surgeries (THR or TKR). These terms are subsequently used throughout this protocol.

A similar observational study is currently being conducted by the US Food and Drug Administration (FDA) Sentinel Initiative. The study described in this protocol uses a similar outcome definition as the Sentinel study but differs by other key design and methodological approaches, for example, by separately defining exposure cohorts for each population indicated for DOACs and warfarin. An overview of the Sentinel study is publicly available at https://www.sentinelinitiative.org/drugs/assessments/severe-uterine-bleed-following-novel-oral-anticoagulants-use-propensity-score.

6. STUDY OBJECTIVES

6.1. Primary Objective(s)

The primary research questions include:

1. What is **the incidence of SUB** following exposure to individual DOACs (rivaroxaban, apixaban, dabigatran) and warfarin among women with prior diagnoses for DOAC and warfarin indications (NVAF, VTE, THR or TKR)?
2. Are individual DOACs (rivaroxaban, apixaban, dabigatran) associated with **risk of SUB** compared with warfarin among women with prior diagnoses for DOAC and warfarin indications (NVAF, VTE, THR or TKR)?
3. Are individual DOACs (rivaroxaban, apixaban, dabigatran) associated with **risk of SUB** compared with other individual DOACs among women with prior diagnoses for DOAC and warfarin indications (NVAF, VTE, THR or TKR)?

6.2. Secondary Objective(s)

None.

7. RESEARCH METHODS

7.1. Study Design and Setting

The proposed study will involve clinical characterization and population-level effect estimation using a retrospective new user comparative cohort design among adult women aged 18 years and older who are newly exposed to DOACs or warfarin in a nationally representative population of insured patients in the United States (US). Eligibility criteria for patients in the study target and comparator cohorts are defined in Section 7.3. As the primary purpose of this study is to evaluate the safety event of SUB for DOACs (dabigatran initial approval Oct 2010), the study period will be from 19-Oct-2010 to 31-Dec-2018.

7.2. Describe Data Source(s)

The study will be executed against the following 4 US administrative claims databases. All 4 databases have been standardized into the Observational Medical Outcomes Partnership

(OMOP) Common Data Model (CDM)[a], which includes a standard representation of health care experiences (such as information related to drug utilization and condition occurrence) as well as common vocabularies for coding clinical concepts and enables consistent application of analyses across multiple disparate data[5]. Complete specifications for the extract, transform, and load (ETL) process for each database is available at: https://github.com/OHDSI/ETL-CDMBuilder. These 4 databases were selected for use in this comparative evaluation based on the observed sample of DOAC users and preliminary feasibility assessment indicating reasonable counts of severe uterine bleeding, and their capture of patients across the age spectrum. We assessed the feasibility and ultimately excluded an electronic health record (EHR) data source in the study because it contained insufficient information for constructing accurate on-treatment follow-up time.

All analyses will be performed independently within each of these 4 databases to produce 4 source-specific results for each analysis. No patient-level data will be pooled across the databases for any analysis, in part to preserve internal validity of the comparative analyses within each database and avoid the potential risk of 'double-counting' cases for duplicate patients. Instead, meta-analysis estimates will be computed based on per-database aggregate statistics (See Section 9.3.2.).

7.2.1. IBM MarketScan® Commercial Database (v1103)

IBM MarketScan® Commercial Database (CCAE) represent data from individuals enrolled in United States employer-sponsored insurance health plans. The data includes adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses, and dependents. Additionally, it captures laboratory tests for a subset of the covered lives. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans.

The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data also contain selected laboratory test results (those sent to a contracted thirds-party laboratory service provider) for a non-random sample of the population (coded with LOINC codes).

7.2.2. IBM MarketScan® Medicare Supplemental Database (v1104)

IBM MarketScan® Medicare Supplemental Database (MDCR) represents health services of retirees in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. These data include adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy). Additionally, it captures laboratory tests for a subset of the covered lives.

The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide

[a] https://github.com/OHDSI/CommonDataModel

procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data also contain selected laboratory test results (those sent to a contracted thirds-party laboratory service provider) for a non-random sample of the population (coded with LOINC codes).

7.2.3. IBM MarketScan® Multi-State Medicaid Database (v1105)

IBM MarketScan® Multi-State Medicaid Database (MDCD) adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity and Medicare eligibility. Members maintain their same identifier even if they leave the system for a brief period however the dataset lacks lab result data.

The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data does not contain laboratory results.

7.2.4. Optum© De-Identified Clinformatics® Data Mart – Date of Death (v1107)

Optum© De-Identified Clinformatics® Data Mart Database (OptumInsight, Eden Prairie, MN) is an adjudicated administrative health claims database for members with private health insurance, who are fully insured in commercial plans or in administrative services only (ASOs), Legacy Medicare Choice Lives (prior to January 2006), and Medicare Advantage (Medicare Advantage Prescription Drug coverage starting January 2006). The population is primarily representative of US commercial claims patients (0-65 years old) with some Medicare (65+ years old) however ages are capped at 90 years. It includes data captured from administrative claims processed from inpatient and outpatient medical services and prescriptions as dispensed, as well as results for outpatient lab tests processed by large national lab vendors who participate in data exchange with Optum. This dataset also provides date of death (month and year only) for members with both medical and pharmacy coverage from the Social Security Death Master File (however after 2011 reporting frequency changed due to changes in reporting requirements) and location information for patients is at the US state level. Family identifiers are provided and utilized to infer mother-child linkages.

The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data also contain selected laboratory test results (those sent to a contracted thirds-party laboratory service provider) for a non-random sample of the population (coded with LOINC codes).

7.3. Study Population(s)

Cohorts of mutually exclusive female rivaroxaban, apixaban, dabigatran, and warfarin new users of >=18 years of age with >=183 days of prior continuous database observation were constructed

during the study period between 19-Oct-2010 and 31-Dec-2018. The date of new use in each cohort is referred to as the index date. Patients were excluded for edoxaban exposure, exposure to other exposures of interest (e.g. DOACs for warfarin exposure cohort), hysterectomy, vaginal bleed, and medical, surgical, or transfusion management for vaginal bleeding any time prior to the index exposure. From these base exposure populations, additional inclusion criteria were added to construct new user cohorts with the following prior indications:

- **NVAF diagnosis any time prior** and no VTE (DVT or PE) in the past 183 days and no knee or hip replacement surgery in the past 35 days
- **VTE (DVT or PE) in the past 183 days** and no NVAF diagnosis any time prior and no knee or hip replacement surgery in the past 35 days
- **THR or TKR in the past 35 days** and no non-valvular atrial fibrillation any time prior and no venous thromboembolism (deep vein thrombosis of pulmonary embolism) in the past 183 days

Note all inclusion criteria include the index date (e.g. "any time prior" means all prior observed time including the index date). Patients are considered no longer exposed to the index treatment and to have exited the cohort after the end of the inferred persistent drug exposure, in which persistence is defined as sequential dispensing records where the start of the latter record is within a defined number of days after the end of the former record's days' supply (see Section 9.1).

The concept set expressions and mapped codes for all exposures, conditions, and procedures used to construct the exposure cohorts is available in Annex A. An example of the fully specified, human-readable cohort definition for rivaroxaban new users with prior non-valvular atrial fibrillation is available in Annex B. Therefore 12 new user cohorts will be constructed for the primary objective:

Table 1: Female DOAC and warfarin new user cohorts

Exposure Cohort Name	Cohort Definition Full Specification
Rivaroxaban new users with prior NVAF	https://epi.jnj.com/atlas/#/cohortdefinition/11400
Apixaban new users with prior NVAF	https://epi.jnj.com/atlas/#/cohortdefinition/11401
Dabigatran new users with prior NVAF	https://epi.jnj.com/atlas/#/cohortdefinition/11402
Warfarin new users with prior NVAF	https://epi.jnj.com/atlas/#/cohortdefinition/11403
Rivaroxaban new users with prior VTE	https://epi.jnj.com/atlas/#/cohortdefinition/11404
Apixaban new users with prior VTE	https://epi.jnj.com/atlas/#/cohortdefinition/11405
Dabigatran new users with prior VTE	https://epi.jnj.com/atlas/#/cohortdefinition/11406
Warfarin new users with prior VTE	https://epi.jnj.com/atlas/#/cohortdefinition/11407
Rivaroxaban new users with prior THR OR TKR	https://epi.jnj.com/atlas/#/cohortdefinition/11412
Apixaban new users with prior THR OR TKR	https://epi.jnj.com/atlas/#/cohortdefinition/11413
Dabigatran new users with prior THR OR TKR	https://epi.jnj.com/atlas/#/cohortdefinition/11414
Warfarin new users with prior THR OR TKR	https://epi.jnj.com/atlas/#/cohortdefinition/11415

NVAF: non-valvular atrial fibrillation, VTE: venous thromboembolism, THR: total hip replacement, TKR: total knee replacement

All exposure cohorts will be compared pairwise by prior indication group. A total of 18 comparisons performed for the primary objective are shown in Table 2 below.

Table 2: Primary Objective Target vs. Comparator Cohort Comparisons

Target cohort name	Comparator cohort name
With prior NVAF	
Rivaroxaban new users	Warfarin new users
Apixaban new users	Warfarin new users
Dabigatran new users	Warfarin new users
Rivaroxaban new users	Apixaban new users
Rivaroxaban new users	Dabigatran new users
Apixaban new users	Dabigatran new users
With prior VTE	
Rivaroxaban new users	Warfarin new users
Apixaban new users	Warfarin new users
Dabigatran new users	Warfarin new users
Rivaroxaban new users	Apixaban new users
Rivaroxaban new users	Dabigatran new users
Apixaban new users	Dabigatran new users
With prior THR or TKR	
Rivaroxaban new users	Warfarin new users
Apixaban new users	Warfarin new users
Dabigatran new users	Warfarin new users
Rivaroxaban new users	Apixaban new users
Rivaroxaban new users	Dabigatran new users
Apixaban new users	Dabigatran new users

NVAF: non-valvular atrial fibrillation, VTE: venous thromboembolism, THR: total hip replacement, TKR: total knee replacement

7.3.1. Treatment Group

As described in Section 7.3., all exposure cohorts will be compared pairwise by prior indication group. Treatment groups include the following cohorts:

- Rivaroxaban new users with prior NVAF/VTE/THR or TKR
- Apixaban new users with prior NVAF/VTE/THR or TKR
- Dabigatran new users with prior NVAF NVAF/VTE/THR or TKR

7.3.2. Comparator Group

As described in Section 7.3., all exposure cohorts will be compared pairwise by prior indication group. Comparator groups include the following cohorts:

- Warfarin new users with prior NVAF/VTE/THR or TKR
- Apixaban new users with prior NVAF/VTE/THR or TKR
- Dabigatran new users with prior NVAF NVAF/VTE/THR or TKR

7.4. Outcome(s) of Interest

Patients with severe uterine bleed are defined by having a vaginal bleeding record with a blood transfusion recorded on the same day or a record of surgical management on or within the

subsequent 60 days. The technical definition is below. The concept set expressions and mapped codes for all conditions, procedures, observations, and devices used to construct the outcome cohort is available in Annex A. The definition of severe uterine bleeding as the composite of vaginal bleeding with either transfusion or surgical management is consistent with the Sentinel study (see Section 5). However, the definition in this study combines patients with transfusion or surgical management in a single outcome definition because of insufficient study power for the transfusion managed outcome given exposure cohort restrictions to indication groups. The Sentinel study assesses both a transfusion managed outcome and a surgical managed outcome. The code sets used for vaginal bleeding, transfusion, and surgical management are consistent with those used in the Sentinel study.

Severe uterine bleed with transfusion management on the same day or surgical management on or during the subsequent 60 days:

https://epi.jnj.com/atlas/#/cohortdefinition/14347

Initial Event Cohort
People having any of the following:

- a condition occurrence of [631] Vaginal bleed
    - occurrence start is between 2010-10-19 and 2018-12-31 (inclusive)
    - gender is any of: FEMALE

with continuous observation of at least 0 days prior and 0 days after event index date and limit initial events to: *all events per person.*

Inclusion Rules
Inclusion Criteria #1: Surgical management within 60 days or transfusion management on the same day
Having any the following criteria:

- at least 1 occurrences of a procedure of [631] Transfusion management of vaginal bleed, procedures[3] where event starts between 0 days Before and 0 days After index start date
- or at least 1 occurrences of a device exposure of [631] Transfusion management of vaginal bleed, procedures[3] where event starts between 0 days Before and 0 days After index start date
- Or having [object Object] of the following criteria:
    - at least 1 occurrences of a condition occurrence of [631] Surgical management of vaginal bleed, conditions and procedures - OPTIMIZED[2] where event starts between 0 days Before and 60 days After index start date
    - or at least 1 occurrences of a procedure of [631] Surgical management of vaginal bleed, conditions and procedures - OPTIMIZED[2] where event starts between 0 days Before and 60 days After index start date
    - or at least 1 occurrences of an observation of [631] Surgical management of vaginal bleed, conditions and procedures - OPTIMIZED[2] where event starts between 0 days Before and 60 days After index start date
    - Or having [object Object] of the following criteria:
        - at least 1 occurrences of a condition occurrence of [631] Uterine leiomyoma[4] where event starts between 0 days Before and 60 days After index start date
        - and at least 1 occurrences of a procedure of [631] Myomectomy procedures[1] where event starts between 0 days Before and 60 days After index start date

End Date Strategy
Date Offset Exit Criteria
This cohort definition end date will be the index event's start date plus 60 days
Cohort Collapse Strategy:
Collapse cohort by era with a gap size of 60 days.

7.4.1. Negative and positive control outcomes

Negative control outcomes, i.e., outcomes known not to be causally associated with any of the exposure cohorts[6], will be used for empirical calibration[7,8]. Conditions must meet the following requirements to be considered as negative controls[9]: (1) have no drug-condition association in published literature with any study exposure (2) not listed in the "Adverse Drug Reactions" or "Post-marketing" section in the US product label, (3) not considered a FAERS signal, (4) have no indication or contraindication listed in the OMOP Vocabulary for the drug-condition pair, (5) are not general concepts (e.g. finding of trunk), (6) are not considered a drug induced concept, or (7) not considered a pregnancy related concept.

The same analysis that will be performed for each pairwise comparison to assess the risk of SUB (see Sections 9.3.1. and 9.3.2) will also be performed to assess the risk of each negative control outcome. Because the negative control qualifying criteria support the *a priori* assertion of no effect, we assume the true relative risk (RR) for each negative control outcome is 1, and the difference between RR=1 and the observed effect estimate will be considered error, encompassing both random and potentially systematic.

In addition to negative control outcomes, we will also generate and include synthetic positive control outcomes. Positive control outcomes developed on the basis the observed negative controls, but where the true effect size is artificially increased to a desired effect size by injection of additional, simulated outcomes[7]. To preserve the observed confounding structure in the population, these additional outcomes are sampled from predicted probabilities generated using a predictive model fit to identify patients similar to those with observed negative control outcomes during the study time-at-risk. Synthetic positive control outcomes will be differentially inserted among target patients to artificially create positive control outcomes with true relative risks of 1.5, 2, and 4. Using both negative and positive controls, we will fit a systematic error model and with which we will calibrate the hazard ratio, confidence interval, and p-value to account for residual error not addressed by other methods for confounding control (i.e. propensity score adjustment)[7]. The 127 negative control outcomes used in this study are listed in the Table 3 below.

Table 3: 127 Negative control outcomes used for empirical calibration

Negative Control Condition	Negative Control Condition	Negative Control Condition
Abnormal posture	Functional urinary incontinence	Paranoid schizophrenia
Abscess of abdominal wall	Generalized hyperhidrosis	Paresthesia
Acute bronchospasm	Herpes zoster without complication	Patient dependence on care provider
Acute mycoplasmal bronchitis	Hyperglycemia due to type 2 diabetes mellitus	Postoperative hypothyroidism
Acute upper respiratory infection	Hyperparathyroidism	Posttraumatic stress disorder
Acute vaginitis	Idiopathic sleep related non-obstructive alveolar hypoventilation	
Adjustment disorder with depressed mood	Imaging of abdomen abnormal	
Alcohol dependence	Imaging of gastrointestinal tract abnormal	Pseudobulbar affect
Altered sensation of skin	Impacted cerumen	Psychoactive substance dependence
Anesthesia of skin	Impaired fasting glycaemia	Psychologic conversion disorder
Anhedonia	Impingement syndrome of shoulder region	Pure hyperglyceridemia
Asymptomatic human immunodeficiency virus infection	Incontinence of feces	Restless legs
	Incoordination	Retrograde amnesia
Attention deficit hyperactivity disorder	Increased frequency of urination	Schizoaffective disorder
Bacteremia	Infection by methicillin sensitive Staphylococcus aureus	Schizophreniform disorder
Bacterial infection due to Klebsiella pneumoniae	Infection due to Escherichia coli	Sepsis due to methicillin resistant Staphylococcus aureus
Bipolar I disorder	Ingrowing nail	Severe mixed bipolar I disorder without psychotic features
Bipolar II disorder	Intra-abdominal and pelvic swelling, mass and lump	Severe protein-calorie malnutrition
Blood glucose abnormal	Ketoacidosis in type 1 diabetes mellitus	Shoulder stiff
Borderline personality disorder	Ketoacidosis in type 2 diabetes mellitus	Sickle cell trait
Candida infection of genital region	Knee stiff	Simple goiter
Cannabis abuse	Lichenification and lichen simplex chronicus	Skin sensation disturbance
Cervical somatic dysfunction	Localized enlarged lymph nodes	Slurred speech
Chronic fatigue syndrome	Localized infection of skin AND/OR subcutaneous tissue	Snoring
Chronic hepatitis C	Localized swelling, mass and lump, trunk	Solitary nodule of lung
Chronic pain syndrome	Malnutrition of moderate degree	Spinal stenosis of lumbar region
Chronic post-traumatic stress disorder	Mammography abnormal	Stimulant dependence
Chronic schizoaffective schizophrenia	Methicillin resistant Staphylococcus aureus infection	Suicidal thoughts
Cocaine abuse	Mild protein-calorie malnutrition	Tension-type headache
Continuous opioid dependence	Mixed urinary incontinence	Thyroid function tests abnormal
Coordination problem	Moderate intellectual disability	Thyrotoxicosis without goiter OR other cause
Cramp in limb	Moderate protein-calorie malnutrition	Tobacco dependence syndrome
Crohn's disease	Nasal congestion	Toxic diffuse goiter with no crisis
Dependence on supplemental oxygen	Neurogenic bladder	Type 1 diabetes mellitus uncontrolled
Diabetic ketoacidosis without coma	Nicotine dependence	Type II diabetes mellitus uncontrolled
Difficulty speaking	Non-toxic unimodular goiter	Uncomplicated umbilical hernia
Disorientated	Obsessive-compulsive disorder	Unsteady when standing
Dysphagia	Obstructive sleep apnea syndrome	Urinary incontinence
Eating disorder	Onychomycosis due to dermatophyte	Viral hepatitis B without hepatic coma
Eosinophilic asthma	Open wound of anterior abdominal wall without complication	Viral hepatitis C
Falls	Opioid dependence	Vitamin B deficiency
Female stress incontinence		Vitamin D deficiency
		Wheezing

7.5. Exposure(s) of Interest

See Section 7.3.

7.6. Other Variables of Interest (Demographic Characteristics, Confounders, Effect Modifiers)

Propensity scores (PS) will be used as an analytic strategy to reduce potential confounding due to imbalance between the target and comparator cohorts in baseline covariates. The PS is the probability of a patient being classified in the target cohort vs. the comparator cohort, given a set of observed covariates (see Section 9.3.1).

The types of baseline covariates used to fit the propensity score model will be:

- Demographics
  - Age group (5-year bands)
  - Index year
  - Index month
- Condition occurrence record for the concept or any its descendants observed during 183d on or prior to cohort index
- Drug exposure record for the concept or any its descendants observed during 183d on or prior to cohort index
- Procedure occurrence record for the concept or any its descendants observed during 183d on or prior to cohort index
- Measurement record for the verbatim concept observed during 183d on or prior to cohort index
- Charlson Index - Romano adaptation, using conditions all time on or prior to cohort index
- CHA2DS2-VASc - using conditions all time on or prior to cohort index
- Number of distinct conditions observed in 183d on or prior to cohort index (defined as unique SNOMED condition concepts)
- Number of distinct drugs observed in 183d on or prior to cohort index (defined as unique RxNorm ingredient concepts)
- Number of distinct procedures observed in 183d on or prior to cohort index (defined as unique CPT4/HCPCS/ICD9P/ICD10P concepts)
- Number of distinct observations observed in 183d on or prior to cohort index
- Number of distinct measurements observed in 183d on or prior to cohort index (defined as unique LOINC concepts)
- Number of visits observed in 183d on or prior to cohort index
- Number of inpatient visits observed in 183d on or prior to cohort index
- Number of ER visits observed in 183d on or prior to cohort index

Specific drug exposure concepts that define the target and comparator cohorts will be excluded from the propensity score model fitting. This large-scale empirical adjustment strategy should

address expected confounders, including demographics, outcome risk factors, comorbidities associated with mortality, and health service utilization behavior. The study will be subject to the limitation that some confounders may be unmeasured or inadequately represented in US claims and EHR data, including weight, smoking status, and lifestyle behaviors.

8. SAMPLE SIZE AND STUDY POWER

The sample size of the cohorts is reported in Table 4. These patient counts represent the initial population, prior to statistical adjustment, so provide an upper bound of exposure available for each analysis. For population-level effect estimation, where our aim is to produce an unbiased estimate of the average treatment effect, the precision we will achieve will vary by the incidence rate of each outcome. Because our focus is to estimate the magnitude of the effect, it is acceptable to be underpowered for the analyses, recognizing that this will manifest as wider confidence intervals that account for the random sampling error inherent to the analysis. Smaller sample size for specific comparisons may be associated with larger statistical uncertainty. Small samples may also limit the ability to fit adequate propensity models and thus limit our ability to control confounding. Note that we will not pool the raw data across the 4 databases for analysis. Combining the results from the 4 databases in a meta-analysis (see Section 9.3.2.) will improve precision of the HR estimate as a weighted average of the individual HRs estimated from each database. Hochberg's step-up procedure will be applied to control the family-wise error rate[10] of the pooled results separately for the primary and sensitivity analyses of the primary objectives (i.e. 24 indication comparisons).

There is no *a priori* hypothesis testing for this study, therefore there is no prespecified requirement of sample sizes for the comparative analyses. After all design specifications have been implemented for each pairwise comparison, the minimum detectable hazard ratio will be calculated. The calculation includes a targeted type I error rate (alpha) of 0.05 (2-sided) and a type II error rate (beta) of 0.20 (power=80%) and reports the minimum hazard ratio detectable given the final target and comparator patient count, outcome event count, and TAR[11].

Table 4: Number of female new users of direct oral anticoagulants and warfarin with no prior edoxaban exposure, prior hysterectomy, prior vaginal bleed, and prior medical, surgical, or transfusion management for vaginal bleeding.

Cohort Name	CCAE	MDCR	MDCD	Optum
Female rivaroxaban new users with prior NVAF	7,525	14,817	2,622	21,863
Female apixaban new users with prior NVAF	8,247	18,499	3,363	39,618
Female dabigatran new users with prior NVAF	4,193	8,571	743	7,518
Female warfarin new users with prior NVAF	8,800	28,240	5,530	41,219
Female rivaroxaban new users with prior VTE	12,270	5,834	4,808	14,489
Female apixaban new users with prior VTE	5,203	2,931	2,412	10,020
Female dabigatran new users with prior VTE	326	206	89	434
Female warfarin new users with prior VTE	25,955	17,892	9,284	30,396
Female rivaroxaban new users with prior THR or TKR	18,659	8,216	2,492	14,675
Female apixaban new users with prior THR or TKR	2,077	876	390	2,473
Female dabigatran new users with prior THR or TKR	12	9	<5	11
Female warfarin new users with prior THR or TKR	22,143	12,445	1,948	18,874

NVAF: non-valvular atrial fibrillation, VTE: venous thromboembolism, THR: total hip replacement, TKR: total knee replacement

9. DATA ANALYSIS PLAN

9.1. Calculation of Time-at-Risk

Two time-at-risk (TAR) definitions will be used for follow-up of outcome of interest in this study:

- **Primary TAR**: This on-treatment TAR is defined as the time from 1 day after the cohort start date (which is date of first exposure to the cohort-defining drug) to the end of inferred persistent exposure, the last day of observation, or end of study period, whichever comes first.
  - For DOACs, the final exposure record will allow for no more than a 3-day gap between successive exposure intervals (inferred by days' supply) plus 5 days appended to the last exposure date. This definition is intended to reflect the short half-life (~1 day) of DOACs.
  - For warfarin, the final exposure record will allow for no more than a 7-day gap between successive exposure intervals (inferred using days' supply) plus 5 days appended to the last exposure occurrence of outcome. This definition is intended to reflect the longer half-life (~3-5 days) of warfarin.
- **Sensitivity TAR 1**: This on-treatment TAR is defined as the time from 1 day after the cohort start date (which is date of first exposure to the cohort-defining drug) to the end of inferred persistent exposure, the last day of observation, or end of study period, whichever comes first.
  - For DOACs, the final exposure record will allow for no more than a 30-day gap between successive exposure intervals (inferred by days' supply) plus 5 days appended to the last exposure date. This definition is intended to reflect the dispensing behavior of clinical practice.
  - For warfarin, the final exposure record will allow for no more than a 30-day gap between successive exposure intervals (inferred using days' supply) plus 5 days appended to the last exposure occurrence of outcome. This definition is intended to reflect the dispensing behavior of clinical practice.
- **Sensitivity TAR 2**: The intent-to-treat (ITT) TAR is defined as the time from 1 day after the cohort start date until occurrence of outcome, last day of observation, or end of study period, whichever comes first.

Both TAR definitions will include the requirement that patients have at least 1 day of TAR following exposure to the cohort-defining drug of interest in any pairwise comparison. (See Annex D for on-treatment follow-up time distributions for definitions using 3, 7, 15, and 30-day allowable gaps for inferred persistent exposure).

9.2. Patient Characteristics Summary

A descriptive characterization of patients included in each exposure cohort will be provided for demographics, prior health services utilization, diagnoses, medications, lab tests, and procedures (183d prior to index date), and number and type of prior treatments received. Continuous variables will be summarized using mean (± standard deviation) and median. Counts and proportions will be used to summarize categorical variables. Clinical characterization results will be reported in covariate balance tables for the target and comparator cohort in each pairwise comparison. Covariate balance between the comparison cohorts will be summarized by showing the proportions

and mean values for all baseline covariates with the associated standardized mean difference computed for each covariate. Attrition tables will report the loss of patients from the original target and comparator cohorts to the subpopulations that remain after all design considerations have been applied.

9.2.1. Incidence Analysis

The unadjusted incidence of severe uterine bleeding with surgical management and the incidence of severe uterine bleeding with transfusion management will be calculated for all exposure cohorts during two time-at-risk periods (see Section 9.1) to establish the base rate of event occurrence which will provide context to the subsequent population-level effect estimates. The number of persons, number of events during the time-at-risk period, the incidence proportion per 1000 persons, and the incidence rate per 1000 person-years will be computed for both outcomes, during both time-at-risk periods (see Section 9.1) for all 16 exposure cohorts across all 4 databases. The incidence analyses involve direct observation of the experience of patients, which can provide context about the real-world patterns of event occurrence in different populations but cannot be used for causal inference or to draw comparative conclusions about the effects of any treatment.

9.3. Model Specification

9.3.1. Propensity Score Model Specification

Propensity scores (PS) will be used as an analytic strategy to reduce potential confounding due to imbalance of patient characteristics at baseline between the target and comparator cohort in a pairwise comparison. The PS is the probability of a patient being classified in the target cohort vs. the comparator cohort, given a set of observed covariates. The PS will be estimated for each patient using the predicted probability from a regularized logistic regression model, fit with a Laplace prior (LASSO) and the regularization hyperparameter selected by optimizing the likelihood in a 10-fold cross validation, using a starting variance of 0.01 and a tolerance of 2e-7. Covariates that occur in fewer than 0.1% of the combined target and comparator cohort in a pairwise comparison will be excluded prior to model fitting. Patients will be matched on the PS by two methods:

- **Primary PS strategy**: 1:1 ratio matching of target to comparator patients
- **Sensitivity PS strategy**: 1:100 variable ratio matching of target to comparator patients

Both approaches will use a greedy matching algorithm by applying a caliper of 0.2 of the standard deviation on the logit scale of the PS distribution.

9.3.2. Outcome Model Specification

A Cox proportional hazards regression model will be used to model the time to the first outcome occurrence for the target group relative to the comparator group while accounting for the PS matching. Estimates of risk will be generated as the empirically calibrated hazard ratios (HR), 95% confidence intervals (CI), and p-values (see Section 7.4.1). The uncalibrated HR, CI, and p-value will also be reported. The number of persons, days amount of time-at-risk, and number of outcome events in each cohort in each pairwise comparison after PS adjustment will also be reported.

Two methods for fitting the outcome model will be implemented, each which aligns with 2 PS adjustment strategies (see Section 9.3.1). For the 1:100 variable ratio matching approach, the Cox proportional hazards regression outcome model will be conditioned on the matched sets. For the 1:1 ratio matching approach, the Cox proportional hazards regression outcome model will be unconditional. Empirical and simulated assessment has shown the latter approach demonstrates improved precision and CI coverage relative to other PS adjustment strategies[12,13].

The total number of database-specific, population-level effect estimates that will be generated by this study is the following: 18 indication comparisons * 2 PS adjustment strategies * 3 TAR specifications * 4 databases = 432 estimates. Note that the total number of estimates generated does not necessarily mean the number of estimates that will be reported. The number of estimates reported depends on diagnostic assessment (see Section 9.4).

Small sample sizes of some exposure cohorts may limit the ability to generate population-level effect estimates for which valid inferences can be made. For example, small exposure cohort sample sizes may limit the ability of the PS adjustment strategy to achieve acceptable covariate balance in a pairwise comparison or in conjunction with outcome event occurrence may be underpowered to detect an estimate of a meaningful magnitude. Rather than deciding a priori to not make certain comparisons on this basis, this study will generate a full set of population-level effect estimation diagnostics, including empirical calibration, for all pre-specified pairwise comparisons; the estimates for target-comparator-outcome-analysis-databases combinations that acceptably pass all study diagnostics will be reported (see Section 9.4.4.). Consistent application of pre-specified methods in high throughput observational studies may reduce results reproducibility problems observed when study design decisions are made on a study- or comparison-specific basis[14].

For each target-comparator-outcome-analysis combination, heterogeneity of the hazards ratios will be estimated across the 4 databases, using I^2 as a metric[15]. If there is sufficient homogeneity across sources (I^2<40%)[16], database-specific estimates will be pooled through random effect meta-analysis using the Hartung-Knapp-Sidik-Jonkman inverse-variance method[17]. Pooled results will include p-values corrected for multiple testing using Hochberg's step-up procedure (see Section 8). Where observed heterogeneity across sources is greater than I^2≥40%, pooled estimates will not be generated.

9.4. Evidence Evaluation

For each population-level effect estimate generated by the study, i.e. each target-comparator-outcome-analysis-database combination, we will report diagnostics to assess its potential for bias and threats to its valid interpretation.

9.4.1. Propensity Score distribution

Once the PS model is fit for each pairwise comparison, the PS distribution for the target and comparator cohort will be plotted to evaluate the comparability, as a proxy for exchangeability, of the two cohorts before matching. The plot will be scaled to the preference score, which normalizes for initial cohort size imbalance. If the proportion of patients in clinical equipoise, i.e. the patients

with a preference score between 0.3 and 0.7[18], is less than 50%, then the estimate will be flagged for careful interpretation given potential differences between the analysis sample and the target population.

9.4.2. Covariate Balance Before and After Propensity score matching

Covariate balance will be evaluated by plotting the standardized mean difference (SMD) of each covariate before against the SMD after propensity score matching. After matching SMDs with values of <0.1 are asserted to indicate negligible group differences[19].

9.4.3. Empirical Null Distribution and Systematic Error Model

As described in Section 7.4.1., the distribution of estimates on negative control outcomes (i.e. the empirical null distribution) describes the residual error of a study specification after confounding control has been implemented. Calibration effect plots for the negative controls and synthetic positive controls will be used the visualize the empirical null distribution for assessment.

9.4.4. Evidence Evaluation Results

All evidence evaluation diagnostic results are available in an interactive, web-based tool available on the JnJ network here: https://sharedshiny.jnj.com/user/jweave17/Epi_680/. Table 5 reports whether each exposure cohort comparison in each database achieves adequate balance on observed covariates, which reduces the risk of potential confounding of the population-level effect estimates produced for each comparison after 1:1 propensity score matching. Covariate balance diagnostic results after 1:100 variable ratio matching are similar. Comparisons in the THR or TKR indication group are similarly imbalanced and estimates will not be reported except for rivaroxaban vs warfarin (CCAE, MDCD, Optum), apixaban vs warfarin (Optum), and rivaroxaban vs apixaban (Optum). Database-specific comparisons for which estimates will not be reported in the NVAF and VTE indication groups are indicated in the table. Database-specific comparisons where estimates are not reported will not contribute patient counts, event counts, or estimates to the pooled analysis as described in Section 9.3.2. Although underpowered estimates from database-specific comparisons will contribute to the pooled analysis, comparisons where no events are observed in one treatment arm cannot be included because a standard error is not generated with the estimate which is needed for meta-analysis.

Table 5: Exposure cohort comparisons by database that achieve (pass) or do not achieve (fail) adequate covariate balance after 1:1 propensity score matching. Population-level effect estimates from database-specific exposure cohort comparisons that achieve adequate covariate balance will be reported.

Target cohort*	Comparator cohort	CCAE	MDCD	MDCR	Optum
With prior NVAF					
Rivaroxaban	Warfarin	pass	fail	pass	pass
Apixaban	Warfarin	pass	pass	pass	pass
Dabigatran	Warfarin	pass	fail	pass	pass
Rivaroxaban	Apixaban	pass	pass	pass	pass
Rivaroxaban	Dabigatran	pass	fail	pass	pass
Apixaban	Dabigatran	pass	fail	pass	pass
With prior VTE					
Rivaroxaban	Warfarin	pass	pass	pass	pass
Apixaban	Warfarin	pass	pass	fail	pass
Dabigatran	Warfarin	fail	fail	fail	fail
Rivaroxaban	Apixaban	pass	pass	pass	pass
Rivaroxaban	Dabigatran	fail	fail	fail	fail
Apixaban	Dabigatran	fail	fail	fail	fail
With prior THR or TKR					
Rivaroxaban	Warfarin	pass	fail	pass	pass
Apixaban	Warfarin	fail	fail	fail	pass
Dabigatran	Warfarin	fail	fail	fail	fail
Rivaroxaban	Apixaban	fail	fail	fail	pass
Rivaroxaban	Dabigatran	fail	fail	fail	fail
Apixaban	Dabigatran	fail	fail	fail	fail

NVAF: non-valvular atrial fibrillation, VTE: venous thromboembolism, THR: total hip replacement, TKR: total knee replacement

*Diagnostic results were generated for exposure populations with prior coronary artery disease or peripheral artery disease despite only warfarin and rivaroxaban being approved for this condition. Diagnostics for all comparisons in the exposure populations with prior coronary artery disease or peripheral artery disease failed, and no results will be reported.

Following study diagnostic assessment, the total number of database-specific, population-level effect estimates that will be reported is the following: [18 indication comparisons * 2 PS adjustment strategies * 3 TAR specifications * 4 databases] – [(6 indication comparisons * 2 PS adjustment strategies * 3 TAR specifications * 4 databases) + (2 indication comparisons * 2 PS adjustment strategies * 3 TAR specifications * 3 databases) + (6 indication comparisons * 2 PS adjustment strategies * 3 TAR specifications * 1 database) = 216 estimates. Given that at least 2 database-specific estimates are required to generate a pooled result per comparison, and assuming across-database heterogeneity is acceptable for all comparisons, the minimum number of primary analysis estimates to be reported is 12 (10 pooled estimates + 2 single database estimates). Assuming across-database heterogeneity is unacceptable for all comparisons, the maximum number of primary analysis estimates to be reported is 36 single database estimates.

10. STRENGTHS AND LIMITATIONS OF THE RESEARCH METHODS

Strengths of the study include:

- Use of large observational datasets that include a nationally representative sample of insured US adults and provide complementary perspectives about treatment utilization and effects within the US population. The outcome under study is rare and it is unlikely that randomized and/or prospective research would have sufficient statistical power for meaningful study.
- Cohort studies allow direct estimation of incidence rates following exposure of interest, and the new-user design can capture early events following treatment exposures while avoiding confounding from previous treatment effects. New use allows for a clear exposure index date designation.
- Propensity score adjustment allows balancing on many baseline potential confounders.
- The use of a set of negative control outcomes allows for estimation of residual bias and unmeasured confounding inherent to the study design and data.

Limitations include:

- Indication and outcome misclassification are a concern in administrative databases because diagnosis codes intended for reimbursement cannot verify patient clinical condition with certainty.
- The study will be subject to the limitation that some confounders may be unmeasured or inadequately represented in claims data, including weight, smoking status, many clinical measurements, and lifestyle behaviors, such as diet and exercise.
- Causality between drug exposure and any given event cannot be drawn for individual cases.
- Adjustment by propensity score may not completely remove bias due to unmeasured or misspecified confounders.

11. PROTECTION OF HUMAN SUBJECTS

The New England Institutional Review Board (IRB) has determined that studies conducted in IBM MarketScan CCAE, MDCR, MDCD, and Optum Extended DOD are exempt from study-specific IRB review, as these studies do not qualify as human subjects research. All patient data are deidentified in this study. At no time during the study will the sponsor receive patient identifying information except when it is required by regulations in case of reporting adverse events.

12. SAFETY DATA COLLECTION AND REPORTING

This study uses coded data that already exist in an electronic database. In this type of database, the minimum criteria for reporting an adverse event (i.e., identifiable patient, identifiable reporter, a suspect product, and event) are not available, and adverse events are not reportable as individual case safety reports. The study results will be assessed for medically important results.

13. PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

The protocol will be registered at www.clinicaltrials.gov after finalization. Results will be reported to the registration location within 12 months of completion. Additionally, results will be submitted for peer-reviewed publication.

14. LIST OF TABLES & FIGURES

Standardized results will be generated adherent to the data model described here: https://github.com/OHDSI/SkeletonComparativeEffectStudy/blob/master/inst/doc/DataModel.pdf . All report artifacts such as tables and figures will generated from this result set.

15. ANNEX (LIST OF STAND-ALONE DOCUMENTS)

Document Number	Date	Title
1	21 Feb 2020	Annex_A_concept_set_expressions.xlsx
2	26 Feb 2020	Annex_B_human_readable_cohort_definitions.doc
3	23 Mar 2020	Annex_C_doac_starting_doses.csv
4	23 Mar 2020	Annex_D_follow_up_time_distribution_3_7_15_30_day_gaps.csv

16. REFERENCES

1. Martinelli I, Lensing AW, Middeldorp S, et al. Recurrent venous thromboembolism and abnormal uterine bleeding with anticoagulant and hormone therapy use. *Blood* 2016; 127(11): 1417-25.
2. Brekelmans MP, Scheres LJ, Bleker SM, et al. Abnormal vaginal bleeding in women with venous thromboembolism treated with apixaban or warfarin. *Thrombosis and haemostasis* 2017; 117(4): 809-15.
3. Scheres L, Brekelmans M, Ageno W, et al. Abnormal vaginal bleeding in women of reproductive age treated with edoxaban or warfarin for venous thromboembolism: a post hoc analysis of the Hokusai-VTE study. *BJOG : an international journal of obstetrics and gynaecology* 2018; 125(12): 1581-9.
4. Huisman MV, Ferreira M, Feuring M, Fraessdorf M, Klok FA. Less abnormal uterine bleeding with dabigatran than warfarin in women treated for acute venous thromboembolism. *Journal of thrombosis and haemostasis : JTH* 2018; 16(9): 1775-8.
5. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association : JAMIA* 2015; 22(3): 553-64.
6. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. *Drug safety* 2013; 36 Suppl 1: S33-47.
7. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences of the United States of America* 2018; 115(11): 2571-7.
8. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in medicine* 2014; 33(2): 209-18.
9. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. *Journal of biomedical informatics* 2017; 66: 72-81.
10. Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* 1988; 75(4): 800-2.
11. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983; 39(2): 499-503.
12. Schuemie MJ CM, Suchard MA, Yang J, Tian Y, Schuler A, Ryan PB, Madigan D, Hripcsak G. How Confident Are We About Observational Findings in Healthcare: A Benchmark Study. *Submitted*.
13. Wang SV, Jin Y, Fireman B, et al. Relative Performance of Propensity Score Matching Strategies for Subgroup Analyses. *American journal of epidemiology* 2018; 187(8): 1799-807.
14. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philosophical transactions Series A, Mathematical, physical, and engineering sciences* 2018; 376(2128).
15. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed)* 2003; 327(7414): 557-60.
16. Cochrane Handbook for Systematic Reviews of Interventions: 9.5.2 Identifying and measuring heterogeneity. Cochrane Reviews; 2011.
17. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC medical research methodology* 2014; 14: 25.
18. Walker A PA, Lauer M, Hornbrook M, Marin M, Platt R, Roger V, Stang P, Schneeweiss S. A tool for assessing the feasibility of comparative effectiveness research. *Comparative Effectiveness Research* 2013; 3: 11-20.
19. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine* 2009; 28(25): 3083-107.