



Applying the OMOP Common Data Model to Facilitate Benefit-Risk Assessments of Medicinal Products Using Real-World Data from Singapore and South Korea

Hui Xing Tan^{1*}, Desmond Chun Hwee Teo^{1*}, Dongyun Lee², Chungsoo Kim³, Jing Wei Neo¹, Cynthia Sung^{1,4}, Haroun Chahed¹, Pei San Ang¹, Doreen Su Yin Tan⁵, Rae Woong Park^{2,3}, Sreemane Raaj Dorajoo¹

¹Vigilance & Compliance Branch, Health Products Regulation Group, Health Sciences Authority, Singapore

²Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea

³Department of Biomedical Sciences, Graduate School of Medicine, Ajou University, Suwon, Korea

⁴Health Services and Systems Research, Duke-NUS Medical School, Singapore

⁵Department of Pharmacy, Khoo Teck Puat Hospital, Singapore

Objectives: The aim of this study was to characterize the benefits of converting Electronic Medical Records (EMRs) to a common data model (CDM) and to assess the potential of CDM-converted data to rapidly generate insights for benefit-risk assessments in post-market regulatory evaluation and decisions. **Methods:** EMRs from January 2013 to December 2016 were mapped onto the Observational Medical Outcomes Partnership-CDM (OMOP-CDM) schema. Vocabulary mappings were applied to convert source data values into OMOP-CDM-endorsed terminologies. Existing analytic codes used in a prior OMOP-CDM drug utilization study were modified to conduct an illustrative analysis of oral anticoagulants used for atrial fibrillation in Singapore and South Korea, resembling a typical benefit-risk assessment. A novel visualization is proposed to represent the comparative effectiveness, safety and utilization of the drugs. **Results:** Over 90% of records were mapped onto the OMOP-CDM. The CDM data structures and analytic code templates simplified the querying of data for the analysis. In total, 2,419 patients from Singapore and South Korea fulfilled the study criteria, the majority of whom were warfarin users. After 3 months of follow-up, differences in cumulative incidence of bleeding and thromboembolic events were observable via the proposed visualization, surfacing insights as to the agent of preference in a given clinical setting, which may meaningfully inform regulatory decision-making. **Conclusions:** While the structure of the OMOP-CDM and its accessory tools facilitate real-world data analysis, extending them to fulfil regulatory analytic purposes in the post-market setting, such as benefit-risk assessments, may require layering on additional analytic tools and visualization techniques.

Keywords: Pharmacovigilance, Health Policy, Risk Assessment, Anticoagulants, Data Visualization

Submitted: October 18, 2021, **Revised:** 1st, February 21, 2022; 2nd, March 28, 2022, **Accepted:** March 30, 2022

Corresponding Author

Sreemane Raaj Dorajoo

Vigilance & Compliance Branch, Health Products Regulation Group, Health Sciences Authority, 11 Biopolis Way #11-01 Helios, Singapore 138667. Tel: +65-6866-1126, E-mail: sreemane_dorajoo@hsa.gov.sg (<https://orcid.org/0000-0002-9613-6994>)

*These authors contributed equally to this study.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

I. Introduction

The changing regulatory landscape of health products has led to an increasing interest in incorporating real-world evidence (RWE) for regulatory decision-making [1]. Regulators are increasingly turning towards analytic frameworks and tools for evidence generation, using real-world data (RWD) to enhance their understanding of the benefits and risks of health products [2]. The key evidentiary needs of regulators include monitoring the effectiveness, safety, and utilization of health products in routine care [3]. Ideally, the evidence generated for regulatory purposes should be scientifically valid, timely, meaningfully contextualized, and sufficient for drawing conclusions while maintaining transparency in the evidence generation process [3].

However, analysing RWD (typically from healthcare databases) and generating RWE that fulfils the aforementioned requirements can be challenging [4]. RWD is predominantly observational in nature and is rarely collected for research purposes. RWD is also often not organized in a form that is suited for analysis. Disparate data coding standards, database architectures, and vocabularies can pose further challenges in generating RWE for informing regulatory decisions, particularly when multiple databases are involved [5]. Using a common data model (CDM) may address some of these challenges by harmonizing the architectures and vocabularies of different databases, which confers analytical interoperability [6]. Converting source data into a CDM creates a copy of the original data and reshapes it to fit the common structure of the CDM. Individual data elements from source are translated to the standardized vocabularies and columns from various source tables are split or merged to fit into target table columns of the CDM [5,7]. CDM-converted databases may then facilitate multi-centre analyses and pooling of results to obtain more robust inferences for various study questions of interest [6,8-10].

While the benefits of CDM conversion for academic purposes are relatively clear, the contribution of CDM conversion towards meeting the broad evidentiary requirements set forth for regulatory purposes remains to be elucidated [3,8]. The aim of this study was to characterize the potential usefulness of CDM conversion by conducting a sample benefit-risk assessment involving CDM-converted data. The Observational Medical Outcomes Partnership (OMOP)-CDM was selected for this study because of its large active user community and use of open-source software, which facilitates code sharing and peer review [6].

II. Methods

This study was performed in two phases. The first phase involved conversion of Electronic Medical Record (EMR) data from their source files to the OMOP-CDM, while the second phase involved an illustrative benefit-risk assessment of the converted data using available tools and code sets.

1. Phase 1: Conversion of Source Data to OMOP-CDM

1) Source data

EMR data originating from a tertiary acute care hospital in Singapore, which provides a wide range of medical and surgical speciality services, were used in this paper. The data contained information on 258,038 unique patients who visited the hospital between January 2013 and December 2016, and comprised approximately 1.1 million records of medical conditions, 5.2 million transactions of ordered medications, and 15.5 million records of laboratory tests and investigations.

2) Conversion of source data to the OMOP-CDM

A precedent for converting a portion of EMR data from Singapore was previously set [11]. Source data tables were transformed into the OMOP-CDM version 5.3.0 through three key steps. Firstly, the source data were profiled to understand its structure and content. Secondly, source data elements were mapped to a specified target location on the CDM schema, through extract, transform, and load (ETL) operations [12]. This step was facilitated by the “Rabbit-In-a-Hat” software, an open-source tool developed by the Observational Health Data Sciences Initiative (OHDSI) for generating flow diagrams illustrating the movement of data elements from source to target [13] (Figure 1). Lastly, vocabulary mappings were applied to translate the codes and values used in the source data to those used in the CDM.

3) Mapping vocabularies from source to target

The data vocabularies employed included the International Classification of Diseases, 9th and 10th revisions (ICD-9, ICD-10) and Systematic Nomenclature of Medicine Clinical Terms (SNOMED-CT) for diagnosis codes, RxNorm Extension for drugs, and Logical Observation Identifiers Names and Codes (LOINC) for laboratory tests and vitals measurements. In general, ETL was performed if a concept was available in the respective vocabularies and could be mapped via database joins with the OMOP concept table (Figures 2, 3). Further details on mapping of drug exposures, diagnosis codes, and laboratory tests can be found in the Supplement A.

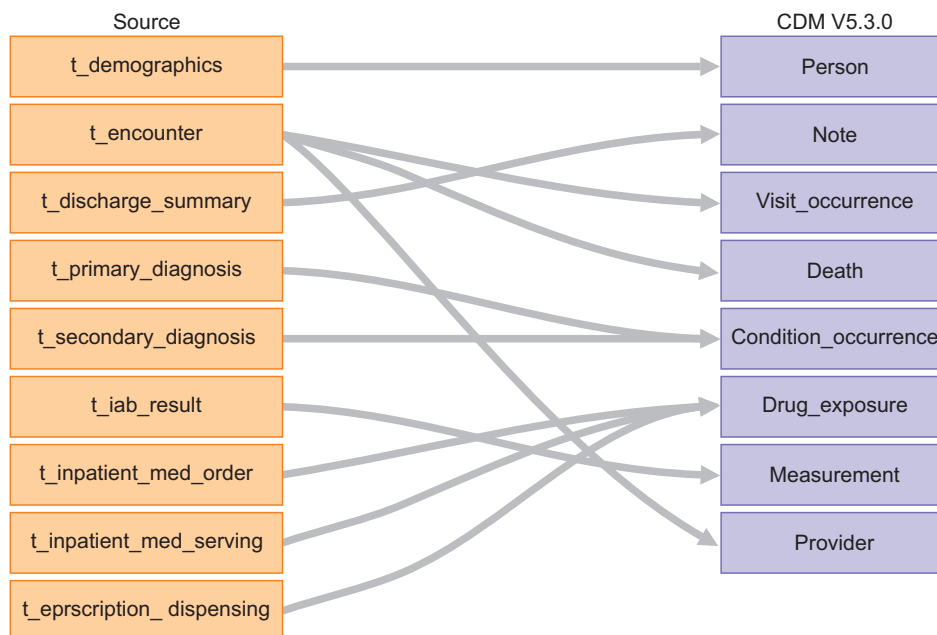


Figure 1. Mapping from source database to target database generated by the Observational Health Data Sciences Initiative (OHDSI) Rabbit-In-a-Hat tool. CDM: common data model.

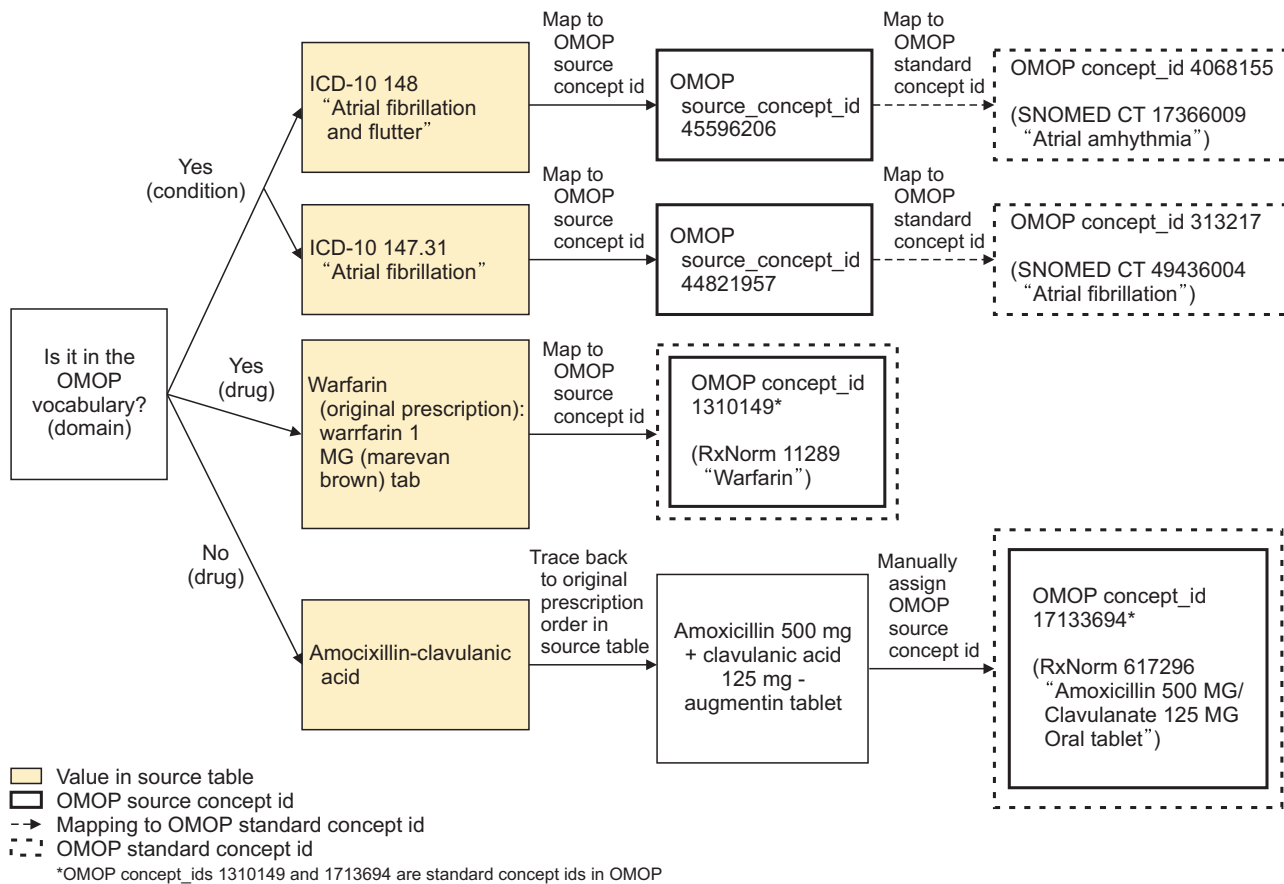
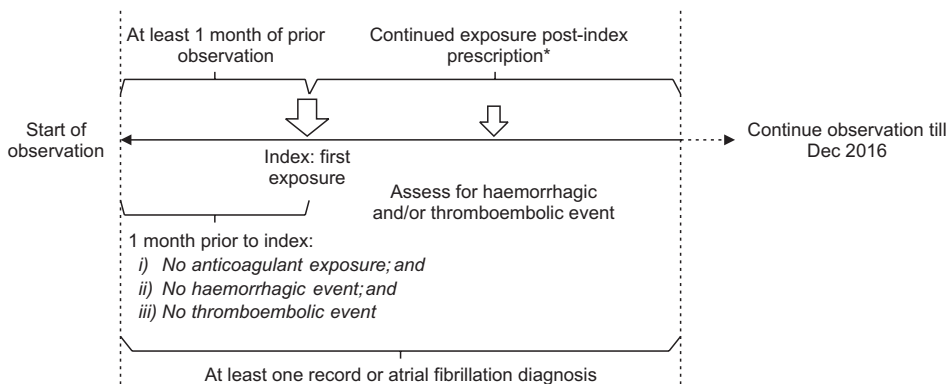


Figure 2. Example of mapping from local concepts to concepts in the Observational Medical Outcomes Partnership (OMOP) vocabulary. ICD-10: International Classification of Diseases 10th edition, SNOMED-CT: Systematic Nomenclature of Medicine Clinical Terms.

Differentiating between source values, source concept ids, and standard concept ids

Source values	Source concept ids	Source concept ids
Source values are fields to maintain the verbatim information from the source database, are stored as unstructured text, and are generally not to be used by any standardized analytics. Type II diabetes mellitus ICD9 code: 250.02 ICD10 code: E11.9	Source concept ids provide a repeatable representation of the source concept, when the source data are drawn from a commonly-used internationally-recognized vocabulary that has been distributed with the OMOP common data model. Concept id1: 44836915 Concept id2: 45561952	Standard concepts ids are those concepts that are used to define the unique meaning of a clinical entity. For each entity there is one standard concept id. 201826

Figure 3. Differentiating between source values, source Concept IDs, and standard Concept IDs. OMOP: Observational Medical Outcomes Partnership, ICD-9: International Classification of Diseases 9th edition, ICD-10: International Classification of Diseases 10th edition.



*At minimum, one prescription for an oral anticoagulant within a 3 months window

Figure 4. Study overview detailing criteria for inclusion and exclusion, exposure, and outcomes.

2. Phase 2: Illustrative Analysis following CDM Conversion

1) Sample cohort assembly and drug exposure

We identified patients diagnosed with atrial fibrillation (AF) without any prior bleeding and/or thromboembolic events for at least 1 month before the first oral anticoagulant (OAC; warfarin or rivaroxaban) exposure in an inpatient or outpatient setting. These patients were followed for at least 3 months after the date of first OAC exposure. Observation ended at the time of bleeding or a thromboembolic event, or at the end of the study. The pre-exposure and follow-up periods were deliberately curtailed because of the limited observation periods available in the data. Patients were included in the final cohort if they had at least one OAC dispensing record in the 3 months following index exposure in an inpatient or outpatient setting. These patients were followed up for the occurrence of bleeding or thromboembolic events at any time after the first OAC exposure. Figure 4 outlines the protocol and definitions applied in this study. Due to the different follow-up times of each patient, a landmark-based analysis at 3 months was performed to equalize the observation times of patients in both arms. More details and phenotype definitions can be found in Supplement A.

2) Visualizing comparative safety, effectiveness, and utilization for benefit-risk assessments

The outcomes of interest in the illustrative analysis were the occurrence of bleeding to represent safety and thromboembolic events to represent effectiveness (or the lack thereof). Patients in the cohort were grouped according to their OAC drug exposure, and only events that occurred during concurrent OAC exposure were extracted.

Adapting a previous OMOP-CDM study by Hripcsak et al. [14], 100%, horizontally stacked, utilization-adjusted bar charts were used to visualize drug utilization (represented by vertical bar thickness) and effectiveness and safety event proportions (represented by horizontal proportion within each bar) to facilitate multiple comparisons in benefit-risk assessments. The charts were created using R version 3.6.0 (<https://cran.r-project.org>). The SQL and R code used in this study is provided in Supplement B.

3) External validation of code on previously converted data and comparisons of the results of the illustrative analysis

An external validation exercise was performed to assess the validity and generalizability of the analytic code on converted OMOP-CDM data. As a mature data partner in the

field of OMOP-CDM, we engaged collaborators from Ajou University, South Korea, who had converted EMR data from Ajou University Medical Center (AUMC)—a 1,200-bed tertiary care facility providing medical and surgical speciality services—into the OMOP-CDM [15,16]. The data from AUMC contained information on about 2,700,000 patients who visited the hospital between January 1994 and December 2020. The code testing exercise and comparison of results were performed to illustrate the potential of generating comparable results from different geographical cohorts of patients and to assess whether any signal of observable differences between agents compared persisted across different cohorts. This study was approved by the Institutional Review Board of Ajou University Hospital (No. AJIRB-MED-MDB-21-191), and the need for informed consent was waived due to the use of de-identified data.

III. Results

1. Phase 1: Conversion of Source Data to OMOP-CDM

Table 1 shows the quantity of data imported in comparison with the source data tables. Over 90% of records from the original table were mapped over to the CDM, except for dispensing records, which included many non-drug items such as foods, syringes, and gauzes. Other types of records not mapped to the CDM included persons with missing birth dates, as well as laboratory records where the corresponding LOINC codes were unavailable or had few records. Diagnoses involving conditional occurrences such as road accidents were excluded, as these were non-crucial for pharmacovigilance studies. Records belonging to 245,561 unique patients were converted into the OMOP-CDM.

2. Phase 2: Illustrative Analysis

A simulated risk-benefit assessment was performed to envisage the potential of OMOP-CDM-converted data in facilitating comparative assessments to inform regulatory decision-making. The results of this analysis are intended for illustrative purposes only and are not meant to be interpreted clinically.

In our sample analysis involving OACs for AF, we identified 364 patients from Singapore and 2,055 patients from South Korea who fulfilled the inclusion/exclusion criteria (Figure 5). Most patients were warfarin users: 73.9% (n = 269) in Singapore and 65.4% (n = 1,345) in South Korea. The patients in the Singaporean cohort were older than those in the South Korean cohort. Among warfarin users, the median (interquartile range) age was 70 years (15 years) in Singapore compared to 63 years (17 years) in South Korea (Table 2). The rivaroxaban users in South Korea tended to be older with median age of 69 years (14 years) than those on warfa-

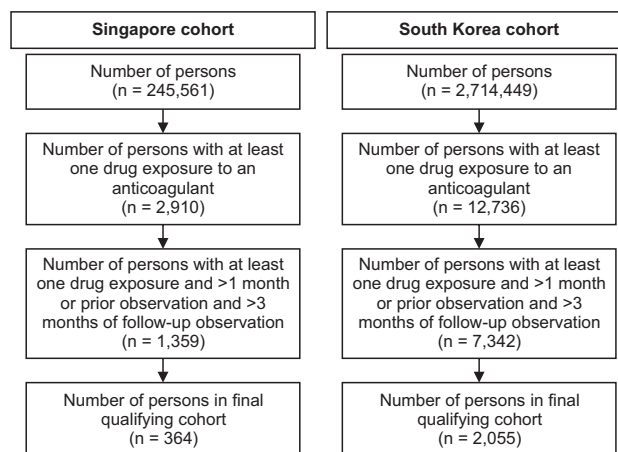


Figure 5. Flow diagram showing the number of persons in the final qualifying cohorts from Singapore and South Korea.

Table 1. Quantity and structure of data imported from a tertiary acute care hospital in Singapore from January 2013 to December 2016

OMOP-CDM table		Source table		
Table name	Number of rows of records	Table name	Number of rows of records	Proportion migrated (%)
person	245,561	t_demographics	258,038	95.2
condition_occurrence	(primary) 210,830	t_primary_diagnosis	222,554	94.7
	(secondary) 799,169	t_secondary_diagnosis	839,265	95.2
measurement	14,116,544	t_lab_result	15,523,576	90.9
visit_occurrence	1,041,587	t_encounter	1,057,263	98.5
drug_exposure	4,378,657	t_eprescription_dispensing ^a	2,147,505	84.8
		t_inpatient_med_order ^b	3,015,159	84.8

^aRefers to outpatient pharmacy orders and inpatient discharge prescriptions.

^bRefers to medications used during inpatient ward stay.

Table 2. Baseline characteristics of the final cohorts from Singapore and South Korea

	Warfarin		Rivaroxaban		Combined		p-value ^d
	Singapore	South Korea	Singapore	South Korea	Singapore	South Korea	
Number of patients	269 (73.9)	1,345 (65.5)	95 (26.1)	710 (34.5)	364 (100)	2,055 (100)	
Age (yr)	70 (15)	63 (17)	71 (15)	69 (14)	72 (15)	66 (17)	<0.001
Sex							<0.001
Male	142 (52.7)	854 (63.5)	44 (46.3)	398 (56.1)	186 (51.1)	1,252 (60.9)	
Female	127 (47.2)	491 (36.5)	51 (53.7)	312 (43.9)	178 (48.9)	803 (39.1)	
Race							<0.001
Korean	NA	1,345 (100)	NA	710 (100)	NA	2,055 (100)	
Chinese	163 (60.6)	NA	66 (69.5)	NA	229 (62.9)	NA	
Malay	66 (24.5)	NA	20 (21.1)	NA	86 (23.6)	NA	
Indian	20 (7.4)	NA	5 (5.3)	NA	25 (6.9)	NA	
Others	20 (7.4)	NA	4 (4.2)	NA	24 (6.6)	NA	
Event outcome ^a							<0.001
Bleeding	81 (30.1)	166 (12.3)	8 (8.4)	47 (6.6)	89 (24.5)	213 (10.4)	
Thromboembolic	32 (11.9)	219 (16.3)	15 (15.8)	64 (9.0)	47 (12.9)	283 (13.8)	
Neither	156 (58.0)	960 (71.4)	72 (75.8)	599 (84.4)	228 (62.6)	1,559 (75.9)	
Concurrent medications (within 7 days before occurrence of bleeding)							NA
Aspirin	7 (2.6)	66 (4.9)	1 (1.1)	2 (0.3)	8 (2.2)	68 (3.3)	
Other NSAIDs ^b	1 (0.4)	7 (0.5)	2 (2.1)	1 (0.1)	3 (0.8)	8 (0.4)	
Clopidogrel	1 (0.4)	15 (1.1)	1 (1.1)	2 (0.3)	2 (0.5)	17 (0.8)	
Other antiplatelets ^c	0 (0)	1 (0.1)	0 (0)	0 (0)	0 (0)	1 (0)	

Values are presented as number (%); for age, the median (interquartile range) are used to indicate the value.

Assignment of patients to drug groupings is based on the latest drug taken by the patient, except in one patient who was on warfarin but who took apixaban for 2 days, and another who was on warfarin but took rivaroxaban for 1 day.

^aBased on the earlier event if patient had records of both bleeding and thromboembolic events.

^bOther non-steroidal anti-inflammatory drugs (NSAIDs) included for analysis were celecoxib, diclofenac, etoricoxib, ibuprofen, indomethacin, ketoprofen, mefenamic acid, meloxicam, naproxen, and piroxicam.

^cOther antiplatelet drugs included for analysis were dipyridamole, eptifibatide, prasugrel, ticagrelor, and ticlopidine.

^dComparing Singapore and South Korea population (using Kruskal-Wallis test for difference in age and Pearson chi-squared test for differences in gender, race, event outcome).

rin. The South Korean cohort also had a noticeable disparity according to sex (60.9% male, 39.1% female), while the Singaporean cohort was more balanced (51.1% male, 48.9% female). The descriptive and clinical characteristics of both cohorts are detailed in Tables 2 and 3.

To visualize the relative proportions of individuals experiencing bleeding (safety) and thromboembolism (effectiveness or lack thereof), while accounting for differences in utilization, we propose the use of 100%, horizontally stacked, bar charts. The left (pink) and right (blue) regions of a bar are used to represent safety and effectiveness, respectively,

while the central region represents the event-free proportion not experiencing any bleeding or thromboembolic events. The sections are coloured to facilitate comparisons within and between agents [17] (Figure 6).

The unadjusted analyses suggested that the overall proportion of bleeding events appeared to be higher among warfarin users than among rivaroxaban users in both cohorts (Figure 6), with the difference being more pronounced in the older Singaporean cohort. However, the higher bleeding risk with warfarin appeared to come at a trade-off for fewer thromboembolic events in the Singaporean cohort.

Table 3. Clinical characteristics of the final cohorts from Singapore and South Korea

	Concept ID	Warfarin		Rivaroxaban	
		Singapore	South Korea	Singapore	South Korea
Number of patients		269 (76.5)	1,345 (65.5)	95 (19.7)	710 (34.5)
Number of diagnoses		310 ^a	1,827	105 ^b	961
Diagnosis (%)					
Atrial flutter	314665	1 (0.4)	0 (0)	0 (0)	0 (0)
Atrial fibrillation	313217	33 (12.3)	0 (0)	4 (4.2)	0 (0)
Atrial arrhythmia ^b	4068155	251 (93.3)	881 (65.5)	92 (96.8)	336 (47.3)
Atrial fibrillation and flutter	4108832	13 (4.8)	0 (0)	2 (2.1)	0 (0)
Atypical atrial flutter	36712986	0 (0)	3 (0.2)	0 (0)	0 (0)
Chronic atrial fibrillation	4141360	0 (0)	71 (5.3)	0 (0)	111 (15.6)
Paroxysmal atrial fibrillation	4154290	0 (0)	772 (57.4)	0 (0)	438 (61.7)
Persistent atrial fibrillation	4232697	0 (0)	68 (5.1)	0 (0)	60 (8.5)
Sick sinus syndrome	4261842	12 (4.5)	30 (2.2)	6 (6.3)	15 (2.1)
Sinus node dysfunction	317302	0 (0)	0 (0)	1 (1.1)	0 (0)
Typical atrial flutter	36714994	0 (0)	2 (0.1)	0 (0)	1 (0.1)
Duration (day)					
Anticoagulant used before occurrence of bleed		336 ± 296	1,501 ± 1,700	295 ± 305	492 ± 534
Anticoagulant used before occurrence of thromboembolic event		369 ± 270	1,654 ± 1,527	243 ± 238	470 ± 436

Values are presented as number (%) or mean ± standard deviation.

^a28 of the 269 patients were co-diagnosed with “atrial arrhythmia” (Concept ID: 4068155) in combination with “atrial fibrillation” (313217) and/or “atrial fibrillation and flutter” (4108832), while nine were co-diagnosed with “atrial arrhythmia” (4068155) and “sick sinus syndrome” (4261842) based on EMR, which is a descendant Concept ID based on OMOP. One patient was diagnosed with “atrial fibrillation” (313217) and “atrial fibrillation and flutter” (4108832) while one patient was diagnosed with “atrial arrhythmia” (4068155), “atrial fibrillation and flutter” (4108832), and “sick sinus syndrome” (4261842).

^bSix of the 95 patients tagged with “atrial arrhythmia” (4068155) were diagnosed with “sick sinus syndrome” (4261842) based on EMR, which is a descendant Concept ID based on OMOP. Three patients were co-diagnosed with “atrial arrhythmia” (4068155) in combination with “atrial fibrillation” (313217) and/or “atrial fibrillation and flutter” (4108832). One patient was diagnosed with “atrial fibrillation” (313217) and “sinus node dysfunction” (317302).

Still, warfarin appeared to have a 3.6-fold higher bleeding risk (30.1% vs. 8.4%, $p < 0.001$), while rivaroxaban had a 1.4-fold higher thromboembolism risk (15.8% vs. 11.9%, $p = 0.331$) in Singapore. If bleeding and thromboembolism are weighted equally in terms of their impact on quality of life and survival, the relative benefits of rivaroxaban appear to outweigh the relative risks of warfarin. This insight is also available when comparing only the event-free proportions of the two agents (grey region), with rivaroxaban having 17.8% fewer overall events (75.8% vs 58.0%, $p = 0.002$) in absolute terms (Figure 6A, 6B). Similarly, in the South Korean cohort, rivaroxaban appears to be the preferred agent because both the proportions of bleeding and thromboembolism were higher among warfarin users than among rivaroxaban users.

However, these comparisons do not consider the different follow-up times of patients receiving the two drugs, which could contribute to the seemingly larger risks, since warfarin had a longer duration of observation in the databases studied. A landmark-based analysis was performed to equalize the observation times of patients in both arms (Figure 6). Interestingly, this eliminated any difference between the two drugs in the South Korean cohort (there were in fact smaller proportions with bleeding and thromboembolic events with warfarin than with rivaroxaban). Similarly, the benefit-risk ratio in the landmark analysis of the Singaporean cohort did not clearly favour one agent over the other (1.0% difference in event-free proportions) (Figure 6C, 6D).

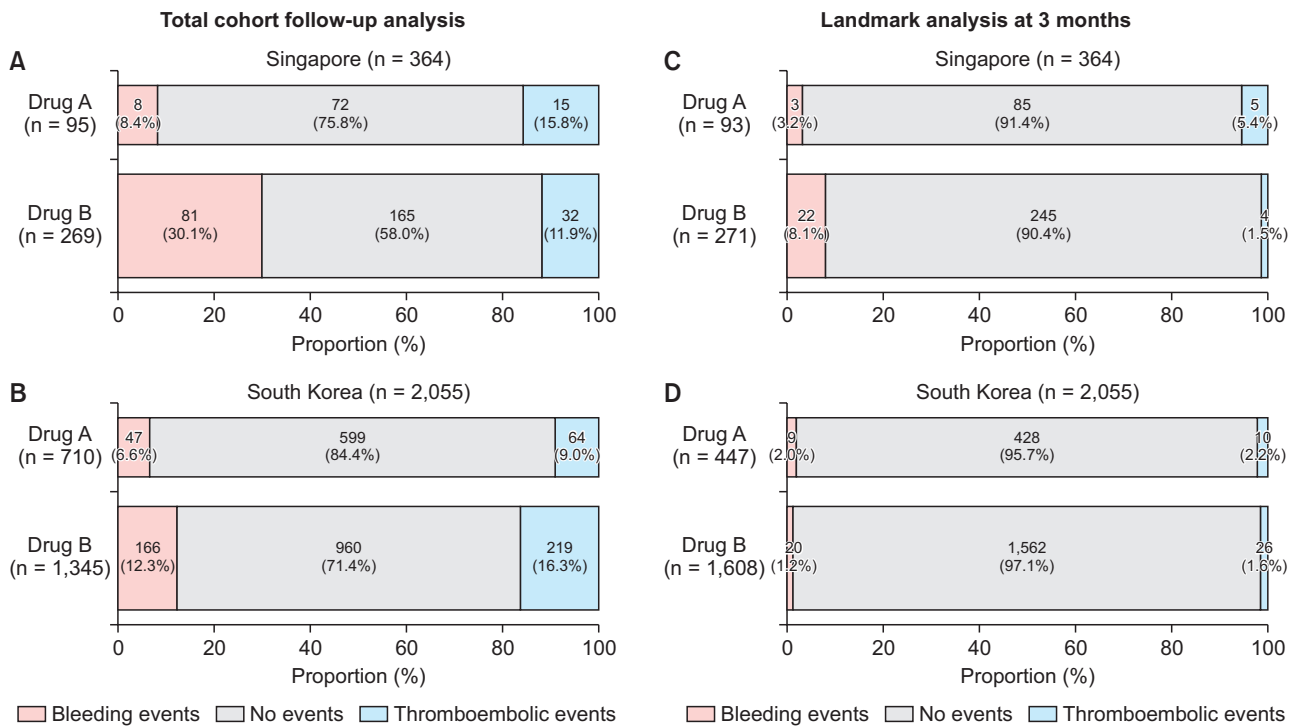


Figure 6. Total cohort follow-up analysis: (A) and (B) are 100%, horizontally-stacked, utilization-adjusted bar charts of effectiveness and safety. The vertical height of each bar is proportional to the number of patients in the Singaporean and South Korean cohorts for 4 years of follow-up. Landmark analysis at 3 months: (C) and (D) are 100%, horizontally-stacked, utilization-adjusted bar charts of effectiveness and safety limited to a follow-up period of three months. The vertical height of each bar is proportional to the number of patients in the Singaporean and South Korean cohorts for 3 months of follow-up. The number of patients experiencing the events of interest are represented as proportions within each bar. Event proportions are unadjusted for confounding factors. Drug A: rivaroxaban, Drug B: warfarin.

IV. Discussion

Our study identified several advantages of converting healthcare databases to the OMOP-CDM related to the conduct of RWD analysis. CDM conversion inevitably involves an inspection of the source data, which can uncover data defects. Tracing to find the root cause of these errors may enable appropriate fixes to be applied. Where unresolvable errors persist, insights as to which sections of the data (or time periods of data) are best left excluded from any analysis are invaluable, as their inclusion may lead to biased results. By exposing data inaccuracies and imposing data cleaning, CDM conversion can also be considered as a process of augmenting source data veracity.

However, CDM conversion alters only the form, but not the substance of the data. This underscores the need to understand the provenance and processes that generated the data and what the data may (and may not) represent. Upon conversion, the set architecture of the CDM, the OHDSI tools, resources and opportunities (i.e., past and ongoing study protocols and, analytic code templates) create a fertile

ecosystem that can speed up analyses, although some modifications and extensions to previously written code are likely required for specific use cases.

Since the previous study by Hripcsak et al. [14] focused on drug utilization patterns in chronic disease management, many code segments were reusable with simple modifications for the purposes of this study. The original code enabled easy specification of the inclusion and exclusion criteria, as well as the observation period of interest. The OMOP-CDM structure contains a derived table (termed the “Drug Era” table) that meaningfully aggregates all drug exposures. This consolidated drug exposure table allows analysts to define and apply the appropriate conditions required for a study (e.g., permitted gap days between prescription fills and stockpiling of previously filled prescriptions). The “Drug Era” table therefore simplifies precise exposure specifications, which are critical in pharmacoepidemiology analyses. Notably, these derived data element features are unavailable in other CDMs, such as the pCORnet, Sentinel, and i2b2 CDMs, which organize medication data at the transaction level, although there may be code segments available to in-

stantaneously aggregate drug exposures during analysis.

The descriptive analysis of OAC usage provides insights on the background incidence of events of interest within a defined observation window. The analysis essentially covers what is described by the US Sentinel Initiative as level 1 analyses [18]. These unadjusted descriptive analyses account for more than 80% of all queries by the US Food and Drug Administration in 2020 to investigate possible drug safety signals. Level 1 analyses help regulators filter signals that warrant subsequent analyses (level 2 and beyond), which typically involve more complex methods for covariate adjustment through various approaches including propensity score matching and stratification [18,19].

Beyond these analyses, comparative assessments may be needed to holistically evaluate the overall impact of any measures undertaken to optimize public health. This would include an analysis of the benefits and risks of a drug relative to that of available alternatives, in the context of its real-world utilization for various therapeutic purposes. Regulatory actions can have far-reaching effects on public health. Benefit-risk assessments facilitate understanding of the potential consequences of various measures undertaken. Large-scale comparative effectiveness analyses have been performed using OMOP-CDM converted data [8-10]. While these are useful, the primary focus and presentation of results in these analyses tend to focus on presenting risks on a relative scale. Regulatory agencies, however, require absolute risk estimates along with real world utilization to establish the net public health impact of policy decisions [3].

To facilitate multiple comparisons as part of benefit-risk assessments, we propose using 100% horizontally stacked bar charts (Figure 6) that amalgamate real-world utilization with effectiveness and safety information. The modularized code provided to derive these charts can be readily extended to other drug classes with composite endpoints to represent outcomes of interest (e.g., major adverse cardiovascular events). The figure facilitates comparisons of the overall prevalence of thromboembolic and bleeding events across anticoagulants at the end of follow-up. Such figures may also be useful for economic analyses, such as cost-effectiveness studies. However, unequal follow-up durations of patients on newer versus older medications are inevitable when using RWD for comparative analyses. To address this issue, we propose applying fixed time-point analyses to eliminate differential time zeros and the potential for immortal-time bias [20] (Figure 6C, 6D).

Our study has a few limitations. Firstly, the CDM conversion was only done using one hospital's data; therefore, any

characterization of the challenges and advantages of conversion may be limited. However, several advantages were identifiable even using only one database. Secondly, an identical analysis was not performed on pre-converted data, as the emphasis was on the possibility of using CDM for regulatory assessments rather than the technical details of conversion. As various data cleaning steps may be undertaken during conversion, not obtaining identical results (pre- and post-conversion) might be an expected outcome. Instead, we validated the analytic code by applying it on an external cohort of patients to indirectly validate the conversion process, while obtaining a separate set of results for comparison [21]. Third, the proposed 100% stacked bar graphs remain an unadjusted descriptive analysis of the rate of events in different populations exposed to comparator agents. Incorporating methods to adjust for confounders and visualize the adjusted event rates would be important areas of future research. Fourthly, the cohorts from the two countries used were demographically different, which could introduce alternative explanations for the study findings; however, studying varied populations may occasionally be desirable to evaluate the consistency of results. Nonetheless, the use of data from two countries and the evaluation of the reproducibility of the analytic code across countries may be seen as a strength of this study, as this demonstrates the potential applicability of this approach to regulators of other countries. Lastly, our study did not evaluate aspects of CDM conversion relating to the mapping coverage and speed relative to other CDMs. These may be of interest to groups looking to embark on the journey of CDM conversion.

Regulatory agencies are increasingly looking to incorporate RWE generated through the analysis of RWD for regulatory decision-making. The findings of this study demonstrate that having access to datasets in the OMOP-CDM format facilitates RWD analysis and can be useful for gleaning insights on comparative drug utilization, effectiveness, and safety for risk-benefit assessments. While the initial conversion is challenging and needs to be done judiciously, the availability of an active community of researchers and open sharing of previously written analytic code promotes transparency and scientific validity in generating RWE that is fit-for-purpose. The ability to refine previously developed analytic code with simple modifications is an important step in harnessing RWD to supplement benefit-risk assessments and enable the conduct of robust evaluations on post-market drug effectiveness and safety use cases, and ultimately make evidence-based decisions to optimise health outcomes.

Conflict of Interest

Rae Woong Park is an editorial member of Healthcare Informatics Research; however, he did not involve in the peer reviewer selection, evaluation, and decision process of this article. Otherwise, no potential conflict of interest relevant to this article was reported.

ORCID

Hui Xing Tan (<https://orcid.org/0000-0003-2284-9788>)
 Desmond Chun Hwee Teo (<https://orcid.org/0000-0001-7641-5462>)
 Dongyun Lee (<https://orcid.org/0000-0002-3678-9862>)
 Chungsoo Kim (<https://orcid.org/0000-0003-1802-1777>)
 Jing Wei Neo (<https://orcid.org/0000-0002-4075-892X>)
 Cynthia Sung (<https://orcid.org/0000-0001-8603-4022>)
 Haroun Chahed (<https://orcid.org/0000-0001-7008-4420>)
 Pei San Ang (<https://orcid.org/0000-0001-7484-5514>)
 Doreen Su Yin Tan (<https://orcid.org/0000-0002-2090-2517>)
 Rae Woong Park (<https://orcid.org/0000-0003-4989-3287>)
 Sreemanee Raaj Dorajoo (<https://orcid.org/0000-0002-9613-6994>)

Supplementary Materials

Supplementary materials can be found via <https://doi.org/10.4258/hir.2022.28.2.112>.

References

- Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther* 2019;105(4):867-77.
- Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Assoc* 2010;17(6):652-62.
- Schneeweiss S, Glynn RJ. Real-world data analytics fit for regulatory decision-making. *Am J Law Med* 2018; 44(2-3):197-217.
- Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet* 2002;359(9302):248-52.
- Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. *EGEMS (Wash DC)* 2017;5(1):14.
- Observational Health Data Sciences and Informatics. The book of OHDSI [Internet]. [place unknown]: Observational Health Data Sciences and Informatics; 2021 [cited at 2022 Apr 15]. Available from: <https://ohdsi.github.io/TheBookOfOhdsi/>.
- Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards implementation of OMOP in a German university hospital consortium. *Appl Clin Inform* 2018;9(1):54-61.
- Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* 2019;394(10211):1816-26.
- You SC, Rho Y, Bickdeli B, Kim J, Siapos A, Weaver J, et al. Association of ticagrelor vs clopidogrel with net adverse clinical events in patients with acute coronary syndrome undergoing percutaneous coronary intervention. *JAMA* 2020;324(16):1640-50.
- Lu Y, Van Zandt M, Liu Y, Li J, Wang X, Chen Y, et al. Analysis of dual combination therapies used in treatment of hypertension in a multinational cohort. *JAMA Netw Open* 2022;5(3):e223877.
- Sathappan S, Jeon YS, Dang TK, Lim SC, Shao YM, Tai ES, et al. Transformation of electronic health records and questionnaire data to OMOP CDM: a feasibility study using SG_T2DM Dataset. *Appl Clin Inform* 2021; 12(4):757-67.
- Observational Health Data Sciences and Informatics. OMOP common data model [Internet]. [place unknown]: Observational Health Data Sciences and Informatics; c2022 [cited at 2022 Apr 15]. Available from: <https://www.ohdsi.org/data-standardization/the-common-data-model/>.
- Observational Health Data Sciences and Informatics. Rabbit-in-a-Hat [Internet]. [place unknown]: Observational Health Data Sciences and Informatics; c2022 [cited at 2022 Apr 15]. Available from: <http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html>.
- Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* 2016;113(27):7329-36.
- Yoon D, Ahn EK, Park MY, Cho SY, Ryan P, Schuemie MJ, et al. Conversion and data quality assessment of electronic health record data at a Korean tertiary teaching hospital to a common data model for distributed network research. *Healthc Inform Res* 2016;22(1):54-8.
- Park RW. Sharing clinical big data while protecting confidentiality and security: Observational Health Data Sci-

- ences and Informatics. *Healthc Inform Res* 2017;23(1): 1-3.
17. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216:574-8.
 18. Sentinel Initiative. Level 1 analyses [Internet]. Silver Spring (MD): US Food and Drug Administration; c2022 [cited at 2022 Apr 15]. Available from: <https://www.sentinelinitiative.org/methods-data-tools/routine-querying-tools/level-1-analyses>.
 19. Sentinel Initiative. Level 2 analyses [Internet]. Silver Spring (MD): US Food and Drug Administration; c2022 [cited at 2022 Apr 15]. Available from: <https://www.sentinelinitiative.org/methods-data-tools/routine-querying-tools/level-2-analyses>.
 20. Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016;79:70-5.
 21. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. *Am J Epidemiol* 2006;163(9):783-9.