Transforming electronic medical records to a common data model for real-world benefit-risk assessments in a tertiary care facility in Singapore

Hui Xing Tan[1], Desmond Teo[1], Haroun Chahed[2], Cynthia Sung[1], Doreen Tan[3], Pei San Ang[1], and Sreemanee Raaj Dorajoo[1]

[1]Health Sciences Authority
[2]Yale-NUS College
[3]Khoo Teck Puat Hospital

September 16, 2020

Abstract

Aim: To assess the feasibility of converting electronic medical records (EMR) into the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) schema and potential for subsequent analyses relating to drug safety. Methods: The EMRs belonging to a tertiary care facility from 2013 to 2016 were mapped onto the OMOP-CDM schema. Vocabulary mappings were applied to translate source data values into OMOP-CDM terminologies. Existing analytic codes from a previous study were modified and applied to conduct an illustrative analysis involving oral anticoagulants (OACs) to mimic analyses that may be part of a typical benefit-risk assessment. A novel visualization is proposed to represent comparative efficacy, safety and utilization in one chart. Results: Records of 245,561 unique patients were mapped onto the OMOP-CDM. The CDM and analytic code templates simplified the data analysis for the illustrative example. Of 132 patients identified, a majority were warfarin users (76.5%), followed by rivaroxaban (19.7%) and apixaban (3.8%). Following six months of follow up, differences in cumulative incidence of bleeding and thromboembolic events were observable. The proposed visualization may facilitate collective evaluation of differences relating to utilization, efficacy and safety of drugs of interest. Conclusion: OMOP-CDM conversion of RWD may be useful for gleaning insights on comparative drug utilization, efficacy and safety for risk-benefit assessments in post-market regulatory settings.

Introduction

Real world data (RWD) analysis of electronic medical records (EMR) for comparative efficacy and safety assessment is challenging.[1] Two key challenges are factors relating to the data itself and the analytical methods applied on the data. Predominantly observational in nature, RWD is rarely collected and organized in a form suitable for analysis. Disparate data coding standards, database schema architectures and vocabularies – observable even within one organization – can hinder accurate analysis.[2] Adopting a common data model (CDM) may address some of these data challenges.

Converting source data into a CDM creates a copy of the original data, reshapes it to fit the CDM structure. Source data elements are translated to the standard vocabularies (e.g. RxNorm for medication data and SNOMED-CT for diagnosis data, as specified by the CDM) and columns from various source data tables are split or merged, to fit into target table columns in the CDM.[3,4] CDM conversion of EMR has several merits – the most obvious of which being the ability to conduct multi-database studies and pool results for obtaining valid inferences to important clinical questions.[5,6]

However, converting data from EMR sources to CDM can be challenging. The value of conversion may not be as apparent to individual institutions as it is to the research community intending to carry out multi-centre analyses. There are however potential benefits of locally analysing a single healthcare system's data – particularly if the system serves a unique patient population. Healthcare providers often struggle to apply new medical findings to their own patients because the evidence regarding efficacy and safety of a given drug would have been generated by studies involving patients with different clinical characteristics and who were studied under highly controlled research settings.[7,8] An introspective analysis of an institution's EMR data in the post-market setting can generate insights that are directly relevant to the patients that the institution serves.

In this study we explored the usefulness of the CDM by converting EMR data from a tertiary care facility in Singapore, into the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM). Potential advantages of converting to the OMOP-CDM include its large and active user community and its emphasis on open source software use and analytic code sharing and peer review. This could address another key challenge of RWD – its appropriate analysis. Upon conversion, we illustrated the possibilities that CDM conversion can potentially offer by delving deeper into a use case involving oral anticoagulants (OAC) in patients with atrial fibrillation.

Methods

This study was carried out in two phases. The first phase involved converting source data from the EMR system to the OMOP-CDM, while the second phase involved an exploratory analysis of the potential and ease of obtaining required results for comparative assessments.

Phase 1: Conversion of source data to OMOP-CDM

1.1 Source data

The EMR data used in this study originated from a tertiary acute care hospital in Singapore, which provides a comprehensive range of medical and surgical speciality services.

The data contained information on 258,038 patients who had visited the hospital between January 2013 and December 2016. Overall, there were approximately 1.1 million records of medical conditions, 5.2 million transactions of ordered medications and 15.5 million entries of laboratory tests and investigations. Among ordered medications, these were further subdivided into dispensed orders (both outpatient and inpatient settings) and inpatient medications (medications administered during inpatient ward stay).

1.2 Conversion of source data to the OMOP-CDM

We transformed the source data into the OMOP-CDM Version 5.3.0. The conversion process from source to CDM consisted of three key steps.

Firstly, the source data was profiled for better understanding of 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. This step was facilitated by the 'Rabbit-In-a-Hat' software, an open-source tool by OHDSI which can be used to generate flow diagrams illustrating the movement of data elements from source to target (Figure 1). Lastly, vocabulary mappings were applied to translate the codes and values used in the source data to that of the CDM (e.g. ICD 10 codes were mapped to SNOMED CT).

Only relevant tables containing information on visits, diagnoses, medication exposures and laboratory tests were converted into the OMOP-CDM.

1.3 Mapping vocabularies from source to target

The data vocabularies employed were the 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[9-11]. In general, ETL was performed if the concept was available in the respective vocabularies and could be mapped via database joins with the OMOP concept table based on "Concept Name". In the upper branch of Figure 2, under the "Conditions" subgroup in the data source, the concept of "J18.9 (Pneumonia, unspecified)" in the International Classification of Diseases, 10threvision (ICD-10) could be mapped to "233604007 (Pneumonia)" in SNOMED CT, which was mapped to the OMOP standard concept identifier "255848".

If the concept did not exist in the OMOP vocabulary, it was mapped through a manual conversion process to an OMOP concept identifier. An example of this process is shown in the lower branch of Figure 2 and explained under '1.4 Drug Exposures'.

1.4 Drug Exposures

To map the source drug data to standard terms in the "Drug Exposure" table of the CDM, the individual drug ingredient names in the source table were manually matched with the "Concept Name" column in the standardized vocabulary (RxNorm).

Some drugs were displayed in the source database as combination products, with separate ingredients represented by their generic drug names. For example, the concept of "Amoxicillin-Clavulanic Acid" did not match directly with a standard drug concept (Figure 2). In this case, the individual active pharmaceutical components of the drug combination were traced back from the prescription order and mapped to the RxNorm code "617296 (Amoxicillin 500 MG / Clavulanate 125 MG Oral Tablet)".

1.5 Diagnosis codes

The hospital migrated from ICD-9 (ICD, 9th revision) to ICD-10 for recording diagnoses during the three-year period for which data were available in this study. Under the OMOP-CDM, both ICD versions could be mapped to the same standard concept ID in the "Condition Occurrence" table. For example, the ICD-9 and ICD-10 codes for "Type 2 Diabetes Mellitus without complications" are "250.02" and "E11.9", respectively. The corresponding source concept IDs are "44836915" and "45561952", respectively. Both of these map to the same OMOP standard concept ID "201826".

Some data cleaning steps were required during the ETL process for diagnosis codes. For example, in the source database, the diagnosis code "I.255 Ischaemic cardiomyopathy" had the decimal place shifted by two places and should have been the ICD-10 code "I25.5". Since both the code and text were in the source database, the proper mapping was rectified by moving the decimal two places to the right and confirming that the text with the corrected code matched the OMOP descriptor. Codes that had more granularity than present in the OMOP vocabulary were mapped to the parent code, such as "E14.69 Unspecified diabetes mellitus" which was mapped to the concept ID corresponding to the parent code of "E14.6. Unspecified diabetes mellitus" in ICD-10.

1.6 Laboratory tests and investigations

In-house codes were used at source for laboratory tests. The hospital provided a mapping for a portion of the internal codes to LOINC codes (used by OMOP-CDM). An open-source mapping coverage was used to transform LOINC codes to the OMOP standard concept IDs in the "Measurements" table. For the remaining laboratory tests, the description of the test and laboratory units were used to bridge the source data to LOINC codes using Athena11Athena is an OHDSI vocabularies repository managed by Odysseus Data Services, Inc. It is hosted as a web application for distributing and browsing standardised vocabularies for all instances of an OMOP-CDM. as a lookup resource. Once this mapping was completed, these LOINC codes were easily mapped to standard concept IDs.

Phase 2: Illustrative analysis following CDM conversion

Following OMOP-CDM conversion, we repurposed relevant components of code written for a previously published OMOP-CDM-based study on treatment pathways by Hripcsak et al.[12] Given that the code had been written for analyzing OMOP-CDM converted data to understand drug utilization patterns in chronic disease management, many code segments were reusable with simple modifications for the purposes of this study.[12]

2.1 Sample cohort assembly and drug exposure

We identified patients diagnosed with atrial fibrillation (AF) without any prior haemorrhagic and/or thromboembolic events for at least three months before the first OAC exposure in an inpatient or outpatient setting. These patients were followed for at least six months from the date of first OAC exposure. Patients were included in the final cohort if they had at least one OAC dispensing record per 90-day period in the six months following index exposure in an inpatient or outpatient setting i.e. a minimum of two drug dispensing records within a six-month period. For patients on warfarin, the presence of an International Normalised Ratio (INR; Concept ID: 3022217) measurement was used as an additional surrogate to ascertain continued warfarin exposure. These patients were followed up for the occurrence of haemorrhagic or thromboembolic events at any time after the first exposure to anticoagulants. Figure 3 outlines the protocol and definitions applied in this study.

The OACs included for analysis were warfarin (Concept ID: 1310149), rivaroxaban (40241331), apixaban (43013024), and dabigatran (45775372). The diagnosis codes for AF included ‘Atrial Flutter’ (Concept ID: 314665), ‘Atrial Fibrillation’ (313217), ‘Atrial Arrhythmia’ (4068155), and ‘Atrial Fibrillation and Flutter’ (4108832). The concept IDs for both thromboembolic and bleeding events are as shown in the Supplementary Table 1.

2.2 Safety outcomes

The primary outcome in the illustrative analysis was the occurrence of any haemorrhagic or thromboembolic event following OAC exposure. These events could occur at any point during the observation period, as long as the criteria for drug exposure was fulfilled (as mentioned above under ‘*2.1 Sample cohort assembly and drug exposure* ’). The list of events considered (along with their OMOP concept IDs) are listed in the Supplementary Material. Two separate queries were written to identify haemorrhagic and thromboembolic outcomes, respectively. Each query returned the sequence of OAC exposure followed by the occurrence of the event(s), if any. Only events that occurred following OAC exposure were extracted. If a patient had both haemorrhagic and thromboembolic events, the earlier outcome was considered.

Patients in the cohort were then grouped according to their OAC drug exposure. Patients in each group were analysed for the type of diagnosis as well as for the use of any concurrent medications that may potentially exacerbate bleed risks, for example anti-platelets like aspirin or clopidogrel, for seven days prior to any haemorrhagic event. Specifically, these concurrent medications were only included if the concurrent period of exposure fell within the preceding seven-day period of the event. For instance, if a patient was dispensed with aspirin for a period of three months on 1st January 2013 and the haemorrhagic event occurred on 1st March 2013, this was considered as concurrent exposure since the dispensing period (and theoretical exposure period) includes the bleeding event date. Conversely, if another patient was dispensed with aspirin for a period of two weeks on 1st January 2013, and the event occurred on 1st February 2013, this would be excluded as concurrent exposure.

2.3 Visualizing comparative efficacy, safety and utilization for benefit-risk assessments

The analytic code adapted from Hripcsak et al had focused their analysis to drug utilization (visualized via sunburst charts), which was insufficient for our purposes of incorporating utilization with efficacy and safety [12]. Taking sunburst charts as inspiration, we iterated over various ways of layering on information on efficacy and safety with utilization. We propose the use of a 100%, horizontally stacked, utilization-adjusted bar chart, that incorporates both drug utilization (represented by vertical bar thickness) as well as efficacy and safety event proportions (represented by horizontal proportion within each bar), to facilitate multiple comparisons in benefit-risk assessments. The chart was created using R, version 3.6.0. SQL and R codes used in this study are provided in the Supplementary Material.

The study was approved by the National Healthcare Group Domain Specific Review Board, with a waiver for informed consent.

Results

Phase 1: Conversion of source data to OMOP-CDM

The source data was mapped onto the OMOP-CDM except for the "Procedures" and "Death" tables, which were absent in the source dataset. Table 1 shows the quantity of data imported and a comparison with source data tables. More than 90% of records from the original table were mapped over to the CDM, except for "Drug Exposure", as dispensing records included many non-drug items such as foods, syringes and gauzes. Other types of records not mapped to the CDM included persons with no information on birth date, and laboratory records where the corresponding LOINC codes were unavailable and with few records. Diagnoses involving condition occurrences such as road accidents were excluded as these were non-crucial for pharmacovigilance studies. Records of 245,561 unique patients were converted into the OMOP-CDM.

Phase 2: Illustrative analysis

In our sample analysis, we identified 132 patients who fulfilled the inclusion/exclusion criteria within the defined study period (Figure 4). Most patients were warfarin users (n=101, 76.5%), followed by rivaroxaban users (n=26,19.7%). There were only 5 patients on apixaban (3.8%) in the period studied (from January 2013 to December 2016). The median age of patients across the different OACs were similar, ranging from 70 to 71 years old. Racial and gender distributions in the cohort were similar to the ratios of the local population in Singapore. The descriptive characteristics of the cohort were detailed in Table 2.

The cohort was analysed for the type of AF diagnosis (Table 3). Majority of patients were assigned the code for "atrial arrhythmia" (92.4% of all patients), followed by "Atrial Fibrillation" (11.4%). There were eight patients in the warfarin arm diagnosed with both "Atrial Fibrillation" and "Atrial Arrhythmia", and four diagnosed with both "Atrial Fibrillation and Flutter" and "Atrial Arrhythmia". Forty patients (30.3%) experienced a haemorrhagic event and 14 (10.6%) experienced a thromboembolic event following OAC exposure over the duration of follow up in our study. Only five patients (3.8%) had a documented exposure of concurrent bleed-inducing medications. All five were on warfarin for anticoagulation.

Overall, the incidence of bleeding was highest among warfarin users at 36.6% (37 out of 101 patients), followed by rivaroxaban at 11.5% (3 out of 26 patients). There were no haemorrhagic events among apixaban users, however there were only five patients in the cohort exposed to apixaban. Thromboembolic events were less common, with eight (7.9%) from the warfarin arm, five (19.2%) from rivaroxaban, and one (20.0%) from apixaban. Majority (59.0%) of patients did not experience any haemorrhagic nor thromboembolic events within the observation period and were 'event-free'.

To visualize relative adverse event rates while accounting for differences in utilization, we extended the previous work by Hripcsak et al and created a novel method of visualizing efficacy, safety and utilization in one chart.[12] Figure 5 shows a 100%, horizontally stacked, bar chart combining relative utilization information (vertical thickness) with the proportions experiencing events of interest (bleeding and thromboembolic events, horizontal proportion). The central grey region represents the proportion not experiencing any haemorrhagic or thromboembolic events at the end of follow up. As patient-time of exposure (and therefore incidence rates) could not be incorporated into the figure, a fixed-timepoint analysis was conducted only including patients who had completed 6 months of continued OAC exposure, to equalize exposure times (Figure 6).

Discussion

This study sought to convert a tertiary institution's EMR database into the OMOP-CDM to assess the value of conversion for post-market regulatory purposes. This is part of Singapore's pioneering effort to assess the feasibility of CDM conversion to obtain insights that help improve drug safety assessments. We find that while data conversion is laborious, there are inherent benefits of undertaking the exercise. CDM conversion is a collaborative effort involving multiple parties such as data scientists and clinical experts that needs to be judiciously undertaken as there are bound to be several data cleaning and editing steps that will transform the data during the conversion. For instance, an appreciation of the coding and medication supply practices and inclinations of an institute can have significant implications to any future analyses. The importance of understanding the provenance and underlying constructs that the data represents cannot be understated in the context of repurposing transactional healthcare data for drawing reliable insights.

Once converted however, the set architecture of the CDM, the OHDSI tools and opportunities available (i.e. past and ongoing study protocols, analytic code templates) may be seen as a fertile ecosystem that can accelerate analyses, although some modifications and extensions to previously written code are likely required for specific use cases.

After making the necessary amendments to the code by Hripcsak et al, we applied it on the OMOP-CDM converted data. The original code allowed us to easily specify the inclusion and exclusion criteria as well as the observation period of interest.[12] Built into the OMOP-CDM is a derived table (termed as the 'Drug Era' table) that aggregates all drug exposures. This consolidated drug exposure table allows analysts to define and apply the appropriate drug exposure conditions required for the study (e.g. permitted gap days in between prescription fills and stockpiling of previously filled prescriptions). The 'Drug Era' table in the OMOP-CDM therefore simplifies precise exposure specification, a key component in any pharmacoepidemiological analysis. Notably, this derived data elements feature is unavailable in other CDMs such as the pCORnet, Sentinel and i2b2 CDMs, which organize medication data at the transaction level [13-16].

In the illustrative (uncontrolled) analysis of AF patients, most (76.5%) were warfarin users, followed by rivaroxaban (19.7%) and apixaban (3.8%) users. This is not surprising given that warfarin had been registered in Singapore since 1995 and would be more commonly used in patients requiring anticoagulation [17]. The DOACs, however, were more recent entrants to the market. Rivaroxaban was registered in 2008, and apixaban in 2012. Since the dataset analysed only included drug dispensing data from 2013 to 2016, this explains the smaller number of patients on rivaroxaban and apixaban. Dabigatran had been registered since 2009. However, we did not find any patients on dabigatran that fulfilled our inclusion criteria.

The three agents compared appear to have differences in event rates for efficacy and safety. Nonetheless, this was meant to be a descriptive analysis only and no adjustment for baseline differences in patient populations are performed. The relative proportions of adverse events among warfarin and DOACs may be somewhat expected given previous findings in clinical trials [18-20], though our study showed higher numerical percentages across all incidences, which may be the effect of clinical trials often selecting healthier populations through stringent selective inclusion/exclusion criteria. This would exclude patients with higher risks of AEs. This study also included a wide variety of ICD codes involving haemorrhagic or thromboembolic events, whereas most studies were confined to ICD codes relating to ischaemic stroke, gastrointestinal bleeds, and intracranial haemorrhage. Additionally, the number of patients using apixaban from our final cohort was very small, hence it was challenging to identify any pattern of usage or trend in events. The higher numerical percentages of events in our study population may also be an indication of real world prevalence of adverse events with OAC in Singapore. [21]

To visualize the results, we extended the work by Hripcsak et al and proposed the use of a 100% horizontally stacked, weighted bar chart (Figure 5) that amalgamates utilization data with efficacy and safety event information to facilitate multiple comparisons between agents and make available the code that we used to derive Figure 5. The figure potentially facilitates comparisons of the overall prevalence of thromboembolic and bleeding events across the agents at the end of follow up. However, the chart does not account for differences in exposure times that would influence the number of patients experiencing the events of interest. Indeed, patients on warfarin had been exposed to warfarin for a longer period of time compared to those on rivaroxaban and apixaban and were therefore more likely to experience events. Therefore, incidence rates could not be represented even though this was a longitudinal study. To resolve the imbalances in exposure times, a fixed time-point analysis was undertaken (Figure 6). Comparing Figures 5 and 6 reveal that while bleeding events occur at any time during follow up, the majority of thromboembolic events appear to occur within 6 months of initiating therapy.

Our study is not without limitations. Firstly, our data was obtained from a limited period and may be insufficient for studying chronic drug exposures and adverse events. The data was from only one hospital, which could skew the findings depending on the usage preferences of OAC in that hospital. It is precisely for this reason that a formal controlled analysis was not undertaken to avoid a potentially biased comparative assessment of the agents, opting instead to perform an illustrative analysis only. To fully leverage the potential benefits of a CDM and provide a more accurate overview of patient journeys, EMR data from more healthcare institutions could be included into future analyses. The application of propensity score matching prior to generating the weighted bar charts could equalize the cohorts and their risk factors to render comparable charts, but this was not viable given that there were few patients at the outset. INR was used as a surrogate measure for duration of drug exposure to warfarin. This is a blunt method to correlate drug exposure as it assumes the presence of the INR test as an indication of the patient taking warfarin. Additionally, the data sources could only capture INR tests performed in public healthcare institutions contributing to the EMR database. Point-of-care testing of INR at satellite care sites or at patients' homes were not included and hence the frequency of testing may be under-represented. Furthermore, in comparison, DOACs do not have a similar indirect measure to rely upon.

Nonetheless, the primary purpose of our study was to evaluate the potential value of adopting a CDM. While our proposed bar chart does not account for other confounding factors that may influence haemorrhagic and thromboembolic event rates observed, it provides an overview of the relative utilisation and adverse event incidence in a specified population of interest– an important first step to identifying insights that warrant more rigorous observational studies it. Within the OHDSI community, similar tools such as the Cohort Characterisation tool in ATLAS have been developed, which allows for viewing of incidence rates of selected events in cohorts of interest. Our stacked bar chart is a quick visual that allows for similar comparison of incidence rates, albeit across users of different drug types. The sharing of code written for the same CDM format could also enable other researchers to conduct the same analysis on their own databases.

Conclusion

The findings 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, efficacy and safety, for risk-benefit assessments. While the initial conversion is challenging and needs to be done judiciously, the availability of a community of researchers and the sharing of previously written analytic code which serve as templates makes the process worthwhile. The ability to refine previously developed analytic codes with simple modifications is an important step in harnessing RWD to supplement benefit-risk assessments and creates a perpetual cycle of tool refinement needed to address the challenges of generating valuable insights from RWD and enable the conduct of robust evaluations on post-market drug efficacy and safety use cases, and ultimately make evidence-based decisions to optimise health outcomes.

References

1. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet.* 2002;359(9302):248-252.

2. 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.

3. Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *EGEMS (Wash DC).* 2014;2(1):1110.

4. Maier C, Lang L, Storf H, et al. Towards Implementation of OMOP in a German University Hospital Consortium. *Appl Clin Inform.*2018;9(1):54-61.

5. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform.* 2015;216:574-578.

6. Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis.*Lancet.* 2019;394(10211):1816-1826.

7. Schneeweiss S. Learning from big health care data. *N Engl J Med.* 2014;370(23):2161-2163.

8. Leather DA, Jones R, Woodcock A, Vestbo J, Jacques L, Thomas M. Real-World Data and Randomised Controlled Trials: The Salford Lung Study. *Adv Ther.* 2020;37(3):977-997.

9. International Health Terminology Standards Development Organisation. SNOMED CT and Other Terminologies, Classifications and Code Systems. 2019; https://www.snomed.org/snomed-ct/sct-worldwide. Accessed 8.11.2019.

10. Regenstrief Institute. About LOINC: Logical Observation Identifiers Names and Codes https://www.loinc.org/about Accessed 9.11.2019.

11. National Institute of Health U.S. National Library of Medicine. Unified Medical Language System®: RxNorm Overview https://www.nlm.nih.gov/research/umls/rxnorm/overview.html Accessed 8.11.2019.

12. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A.* 2016;113(27):7329-7336.

13. The Sentinel Initiative. The Sentinel Common Data Model https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model. Accessed 18.04.2020, 2020.

14. OHDSI. The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html. Accessed 14.04.2020, 2020.

15. pCORnet. pCORnET: Common Data Model (CDM) Specification, Version 5.1 https://pcornet.org/wp-content/uploads/2019/09/PCORnet-Common-Data-Model-v51-2019_09_12.pdf. Accessed 19.04.2020, 2020.

16. i2b2 Common Data Model https://www.i2b2.org/software/projects/datarepo/CRC_Design_Doc_13.pdf. Accessed 18.04.2020, 2020.

17. Health Sciences Authority Singapore. Infosearch - Register of Therapeutic Products. . https://www.hsa.gov.sg/e-services/infosearch. Accessed 09.11.2019.

18. Malik AH, Yandrapalli S, Aronow WS, Panza JA, Cooper HA. Meta-Analysis of Direct-Acting Oral Anticoagulants Compared With Warfarin in Patients >75 Years of Age. *Am J Cardiol.*2019;123(12):2051-2057.

19. Connolly SJ, Ezekowitz MD, Yusuf S, et al. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med.*2009;361(12):1139-1151.

20. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med.*2011;365(10):883-891.

21. Wee XT, Ho LM, Ho HK, et al. Incidence of thromboembolic and bleeding events in patients with newly diagnosed nonvalvular atrial fibrillation: An Asian multicenter retrospective cohort study in Singapore. *Clin Cardiol.* 2017;40(12):1218-1226.

Tables

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

OMOP CDM Tables	**OMOP CDM Tables**	**Source Tables**	**Source Tables**	**Source Tables**	
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%

OMOP CDM Tables	**OMOP CDM Tables**	**Source Tables**	**Source Tables**	**Source Tables**	
	Secondary: number 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*	2,147,505	84.8%
			t_inpatient_med_order$^+$	3,015,159	

* Refers to outpatient pharmacy orders and inpatient discharge prescriptions

$^+$ Refers to medications used during inpatient ward stay

Table 2. Baseline characteristics of final cohort (unique n = 132)

	Warfarin (n = 101)	Rivaroxaban (n = 26)	Apixaban (n = 5)	χ^2 (df)	p-value (Significance)
No. of patients (%)	101 (76.5)	26 (19.7)	5 (3.8)		
Median age (IQR*)	71 (14)	71 (21)	70 (13)	0.169 (2)	0.919 (NS)
Gender				0.481 (2)	0.786 (NS)
Male (n, %)	48 (47.5)	14 (53.8)	2 (40.0)		
Female (n, %)	53 (52.5)	12 (46.2)	3 (60.0)		
Race				7.05 (6)	0.316 (NS)
Chinese (n, %)	60 (59.4)	19 (73.1)	2 (40.0)		
Malay (n, %)	23 (22.8)	3 (11.5)	3 (60.0)		
Indian (n, %)	8 (7.9)	1 (3.9)	0 (0)		
Others (n, %)	10 (9.9)	3 (11.5)	0 (0)		
Event Outcome§				9.85 (4)	0.0431 (S)
Bleeding (n, %)	37 (36.6)	3 (11.5)	0 (0)		
Thromboembolic (n, %)	8 (7.9)	5 (19.2)	1 (20.0)		
Neither (n, %)	56 (55.5)	18 (69.3)	4 (80.0)		
Concurrent medications (within 7 days before occurrence of bleeding)				NA	NA
Aspirin (n, %)	2 (2.0)	0 (0)	0 (0)		
Other NSAIDs$^+$ (n, %)	2 (2.0)	0 (0)	0 (0)		

	Warfarin (n = 101)	Rivaroxaban (n = 26)	Apixaban (n = 5)	χ^2 (df)	p-value (Significance)
Clopidogrel (n, %)	1 (1.0)	0 (0)	0 (0)		
Other antiplatelets^{++} (n, %)	0 (0.0)	0 (0)	0 (0)		

Note: 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 two days, and another who was on warfarin but took rivaroxaban for one day.

* Inter-quartile range

$^{+}$ Non-steroidal anti-inflammatory drugs (NSAIDs) included for analysis were “aspirin”, “celecoxib”, “diclofenac”, “etoricoxib”, “ibuprofen”, “indomethacin”, “ketoprofen”, “mefenamic acid”, “meloxicam”, “naproxen”, “piroxicam”

$^{++}$ Antiplatelet drugs included for analysis were “dipyridamole”, “eptifibatide”, “prasugrel”, “ticagrelor”, “ticlopidine”

§ Based on earlier event if patient had records of both bleeding and thromboembolic events

Table 3. Clinical characteristics of the final cohort (unique n = 132)

	Warfarin (n = 101)	Rivaroxaban (n = 26)	Apixaban (n = 5)
No. of patients (%)	101 (76.5)	26 (19.7)	5 (3.8)
No. of diagnoses	113*	26	5
Diagnosis (Concept ID)			
Atrial flutter (%) (314665)	1 (1.0)	0 (0)	0 (0)
Atrial fibrillation (%) (313217)	13 (12.9)	2 (7.7)	0 (0)
Atrial arrhythmia (%) $^{+}$ (4068155)	94 (93.1)	23 (88.5)	5 (100)
Atrial fibrillation and Flutter (%) (4108832)	5 (5.0)	1 (3.8)	0 (0)
Duration of anticoagulant used before occurrence of bleed, mean (s.d.$^{++}$)	211 (197)	232 (71)	N.A.§
Duration of anticoagulant used before occurrence of thromboembolic event, mean (s.d.$^{++}$)	152 (155)	151 (151)	152 (0)

* Eight of the 101 patients were co-diagnosed with “atrial fibrillation” (313217) and “atrial arrhythmia” (4068155), while four were co-diagnosed with “atrial arrhythmia” (4068155) and “atrial fibrillation and flutter” (4108832)

$^{+}$ Two of the 94 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.

++ Standard deviation

§ No patients on apixaban had bleeding events

Figure legends

Figure 1: Mapping from source database to target database generated by OHDSI's Rabbit-In-a-Hat tool

Figure 2: Example of mapping from local concepts to concepts in the OMOP vocabulary

Figure 3: Study overview detailing criteria for inclusion and exclusion, exposure and outcomes

Figure 4: Flow diagram showing number of persons in final qualifying cohort

Figure 5. 100%, horizontally-stacked, utilization-adjusted bar charts of efficacy and safety. The vertical height of each bar is proportional to the number of patients in the cohort for 4 years 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: Apixaban, Drug B: Rivaroxaban, Drug C: Warfarin

Figure 6: Landmark analysis at six months; 100%, horizontally-stacked, utilization-adjusted bar charts of efficacy and safety limited to a follow-up period of six months. The vertical height of each bar is proportional to the number of patients in the cohort for 4 years 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: Apixaban, Drug B: Rivaroxaban, Drug C: Warfarin

Figures

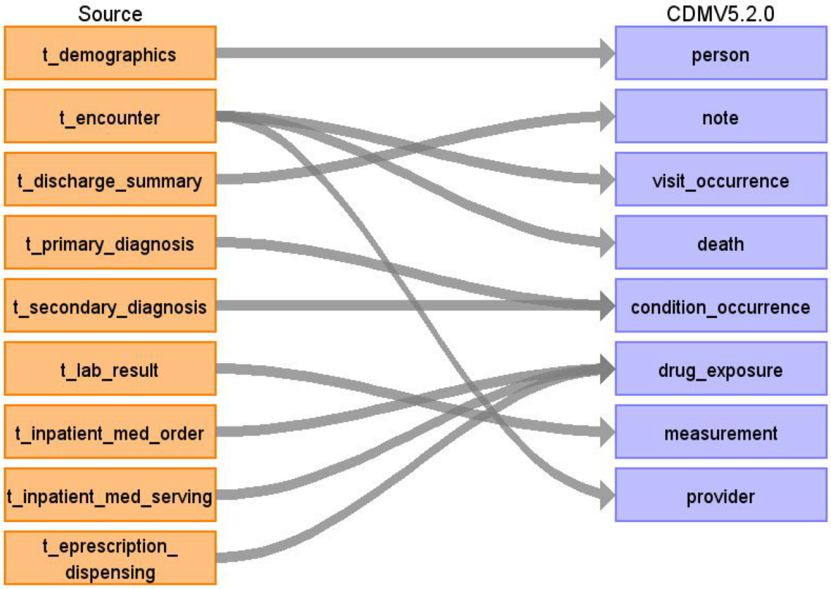
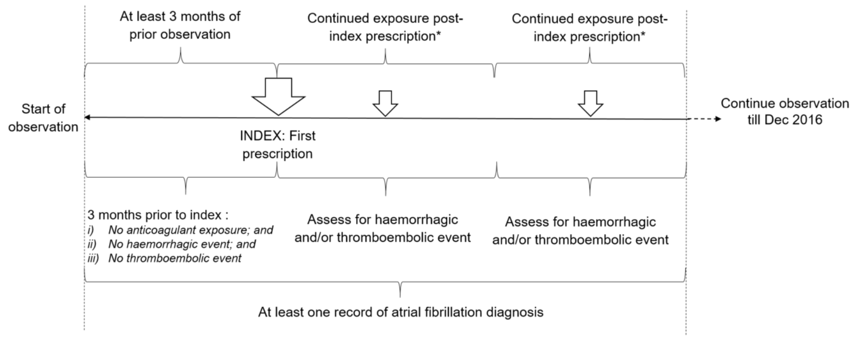


Figure 1: Mapping from source database to target database generated by OHDSI's Rabbit-In-a-Hat tool

Figure 2: Example of mapping from local concepts to concepts in the OMOP vocabulary


At least 3 months of prior observation
Continued exposure post-index prescription*
Continued exposure post-index prescription*
Start of observation
INDEX: First prescription
Continue observation till Dec 2016
3 months prior to index :
i) No anticoagulant exposure; and
ii) No haemorrhagic event; and
iii) No thromboembolic event
Assess for haemorrhagic and/or thromboembolic event
Assess for haemorrhagic and/or thromboembolic event
At least one record of atrial fibrillation diagnosis

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

Figure 3: Study overview detailing criteria for inclusion and exclusion, exposure and outcomes

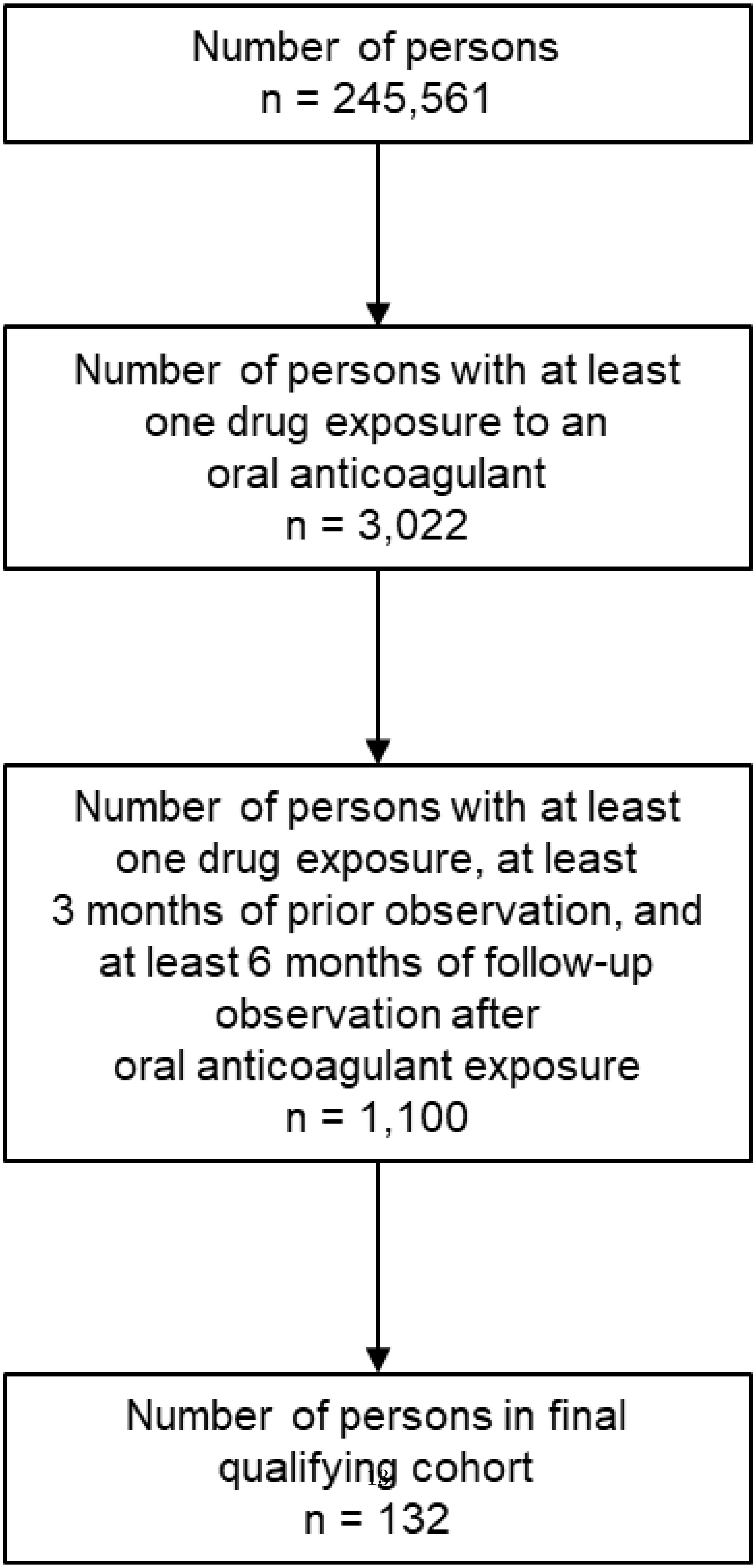

Number of persons
n = 245,561
Number of persons with at least one drug exposure to an oral anticoagulant
n = 3,022
Number of persons with at least one drug exposure, at least 3 months of prior observation, and at least 6 months of follow-up observation after oral anticoagulant exposure
n = 1,100
Number of persons in final qualifying cohort
n = 132

Figure 4: Flow diagram showing number of persons in final qualifying cohort

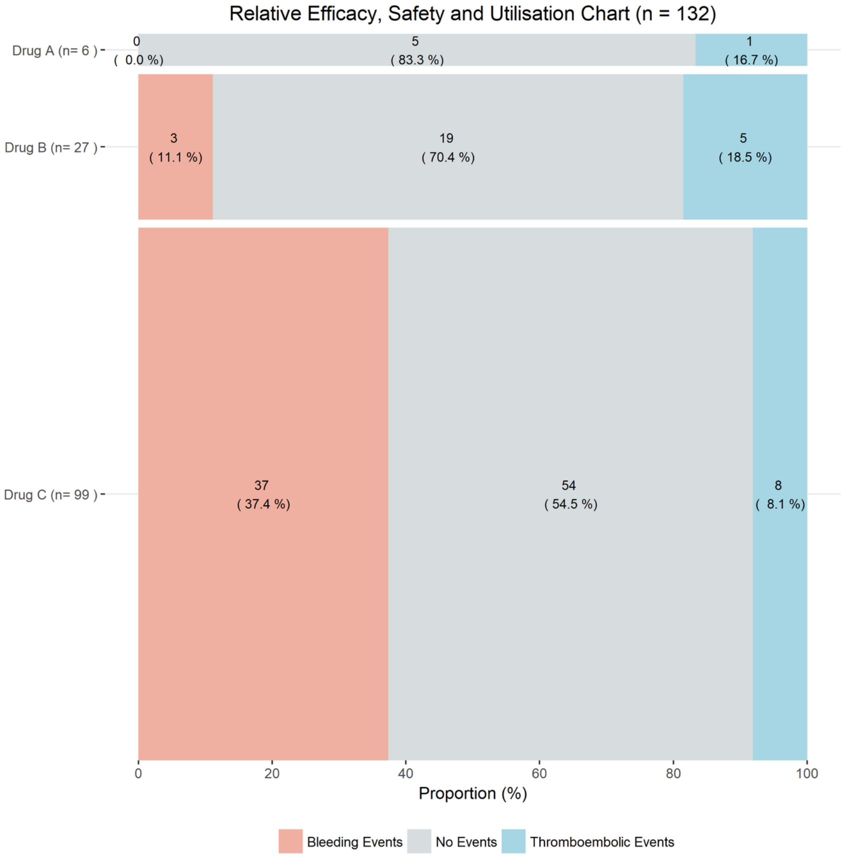


Figure 5. 100%, horizontally-stacked, utilization-adjusted bar chart of efficacy and safety. The vertical height of each bar is proportional to the number of patients in the cohort for 4 years of follow up. The number of patients experiencing the events of interest are represented as proportions within each bar. Event proportions are unadjusted for confounding factors.

Drug A: Apixaban, Drug B: Rivaroxaban, Drug C: Warfarin

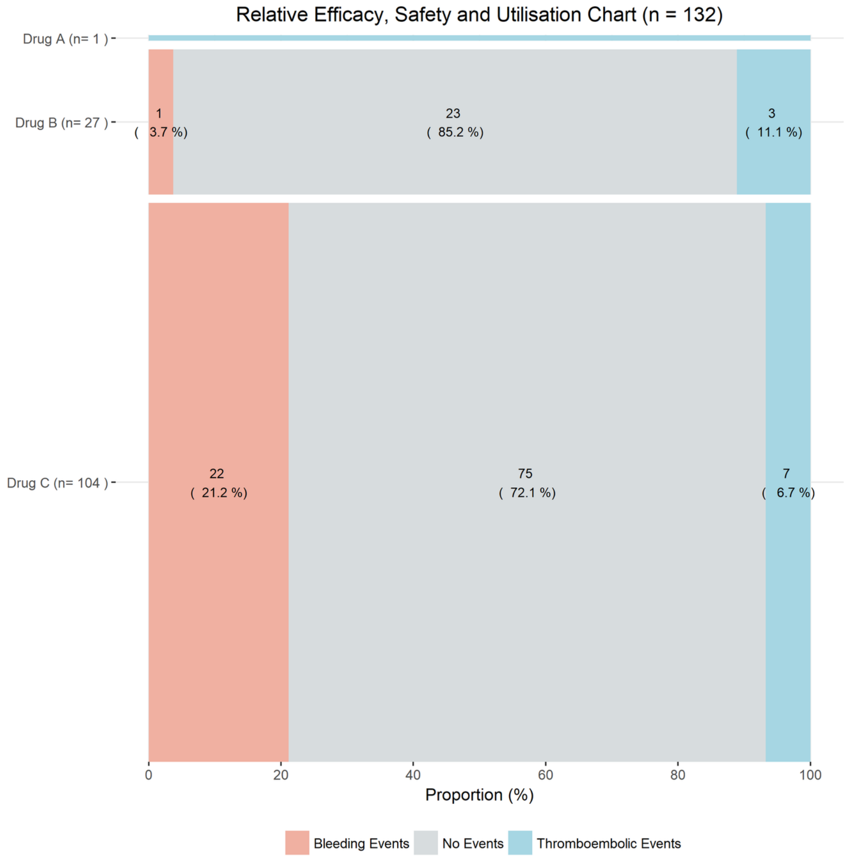


Figure 6: Landmark analysis at six months; 100%, horizontally-stacked, utilization-adjusted bar charts of efficacy and safety limited to a follow-up period of six months. The vertical height of each bar is proportional to the number of patients in the cohort for 4 years of follow up. The number of patients experiencing the events of interest are represented as proportions within each bar. Event proportions are unadjusted for confounding factors

Drug A: Apixaban, Drug B: Rivaroxaban, Drug C: Warfarin