Article published online: 2021-08-11

Transformation of Electronic Health Records and Questionnaire Data to OMOP CDM: A Feasibility Study Using SG_T2DM Dataset

Selva Muthu Kumaran Sathappan[1] Young Seok Jeon[1] Trung Kien Dang[1] Su Chi Lim[2] Yi-Ming Shao[2] E Shyong Tai[3] Mengling Feng[1,4]

[1] Saw Swee Hock School of Public Health, National University Health System and National University of Singapore, Singapore, Singapore
[2] Clinical Research Unit, Khoo Teck Puat Hospital, Singapore, Singapore
[3] Division of Endocrinology, National University Hospital, Singapore, Singapore
[4] Institute of Data Science, National University of Singapore, Singapore, Singapore

Address for correspondence Mengling Feng, PhD, Saw Swee Hock School of Public Health, National University Health System and National University of Singapore, Singapore, Singapore (e-mail: ephfm@nus.edu.sg).

Appl Clin Inform 2021;12:757–767.

Abstract

Background Diabetes mellitus (DM) is an important public health concern in Singapore and places a massive burden on health care spending. Tackling chronic diseases such as DM requires innovative strategies to integrate patients' data from diverse sources and use scientific discovery to inform clinical practice that can help better manage the disease. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) was chosen as the framework for integrating data with disparate formats.

Objective The study aimed to evaluate the feasibility of converting Singapore based data source, comprising of electronic health records (EHR), cognitive and depression assessment questionnaire data to OMOP CDM standard. Additionally, we also validate whether our OMOP CDM instance is fit for the purpose of research by executing a simple treatment pathways study using Atlas, a graphical user interface tool to conduct analysis on OMOP CDM data as a proof of concept.

Methods We used de-identified EHR, cognitive, and depression assessment questionnaires data from a tertiary care hospital in Singapore to convert it to version 5.3.1 of OMOP CDM standard. We evaluate the OMOP CDM conversion by (1) assessing the mapping coverage (that is the percentage of source terms mapped to OMOP CDM standard); (2) local raw dataset versus CDM dataset analysis; and (3) Implementing Harmonized Intrinsic Data Quality Framework using an open-source R package called Data Quality Dashboard.

Results The content coverage of OMOP CDM vocabularies is more than 90% for clinical data, but only around 11% for questionnaire data. The comparison of characteristics between source and target data returned consistent results and our transformed data did not pass 38 (1.4%) out of 2,622 quality checks.

Conclusion Adoption of OMOP CDM at our site demonstrated that EHR data are feasible for standardization with minimal information loss, whereas challenges remain for standardizing cognitive and depression assessment questionnaire data that requires further work.

Keywords
- OMOP CDM
- implementation
- deployment
- electronic health record
- diabetes
- secondary use of EHR data
- cognitive and depression questionnaires
- survey data

received
January 20, 2021
accepted after revision
June 7, 2021

DOI https://doi.org/10.1055/s-0041-1732301.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Background and Significance

Diabetes mellitus (DM) is a major public health problem that affects more than 400 million adults across the globe, and it is projected to surge above 640 million by 2040.[1] It was estimated to be the seventh leading cause of death in 2016, where 1.6 million deaths were due to this medical condition.[2] In Singapore, more than 0.4 million people are affected by this disease, and the number is expected to exceed one million by 2050.[3] Furthermore, it is well known that DM dramatically increases the risk for various complications such as cardiovascular diseases, chronic kidney disease, retinopathy, and foot damage, etc.[4]

However, tackling chronic diseases such as DM is an arduous task and optimal management of it involves a complex interplay of appropriate pharmacological treatments, and a good self-care practice such as adherence to medical therapy, adherence to diet plan, and regular follow-up with health care providers, etc. Additionally, diabetes self-care practice is known to be affected by various other factors such as cognitive dysfunction and depression status of the patients.[5] Therefore, this results in a necessity to collect variety of data about patients such as clinical, lifestyle, cognitive, and depression status to have a holistic understanding of patient profiles to better manage and prevent the disease.

However, having massive volume and variety of data cannot be solely responsible for altering diabetes care unless the framework to turn them into meaningful action are available. In the past decade, several frameworks based on Common Data Model (CDM) standard with well-defined semantics and schema to facilitate data-interoperability have emerged as a solution to address the challenges of integrating multiple data sources to facilitate large-scale health care observational research.[6]

We chose the Observational Medical Outcomes Partnership (OMOP) CDM developed by the Observational Health Data Science and Informatics (OHDSI) community.[7] OMOP CDM was selected over other CDM variants because of its benefits such as (1) exhaustive vocabulary coverage, (2) easy to implement database schema, (3) easy to conduct network research, and (4) active open-source tools community.

Although adoption of OMOP CDM is widespread across the globe,[8–13] it is relatively new in Singapore and health care institutes in the country are not fully aware of the characteristics and benefits of OMOP CDM. Second, majority of the studies published on OMOP CDM reported results only based on EHR, claims, administrative, and registry datasets.[8–13] However, research on application of OMOP CDM for representing data captured through cognitive, and depression assessment questionnaires is very limited and requires further investigation. This leads to the primary objective of our study that is to evaluate the feasibility of converting Singapore-based data source, comprising of EHR, cognitive, and depression assessment questionnaire data to OMOP CDM standard.

While a substantial body of research[8–12] under OMOP CDM has focused on evaluating the OMOP CDM conversion by (1) assessing the mapping coverage (that is the percentage of source terms mapped to OMOP CDM terms); (2) local raw dataset versus CDM-based analysis; and (3) Automated Characterization of Health Information at Large-Scale Longitudinal Evidence Systems (Achilles),[14] a database characterization tool which generates summary statistics about the dataset and performs approximately 160 validation checks on the conformance, completeness, and plausibility of the data in the OMOP CDM. However, in this study, we replace Achilles with Data Quality Dashboard[15] built based on Harmonized Intrinsic Data Quality Framework (HIDQF) that uses a system of categories and contexts that represent strategies for assessing data quality.[16] The benefit of including Data Quality Dashboard method for evaluation is that (1) it provides exhaustive quality checks (compared with Achilles), more than 2500 validation checks based on different dimensions of data quality such as conformance, completeness, and plausibility; and (2) it also offers a dashboard like Achilles but with a precise measure to indicate the quality of the dataset. This can enable data owners across multiple institutions to communicate and compare their data quality results in a well-defined manner.

Objective

The aim of this study is to evaluate the feasibility of converting Singapore-based data source, comprising of EHR, cognitive, and depression assessment questionnaire data to OMOP CDM standard. We evaluate the OMOP CDM conversion by (1) assessing the mapping coverage (that is the percentage of source terms mapped to OMOP CDM terms); (2) local raw dataset versus CDM based analysis; and (3) implementing HIDQF using Data Quality Dashboard.[15]

Additionally, we also validate our OMOP CDM installation by executing a simple treatment pathways study as a proof of concept.

Methods

Dataset

Our source data extracted from the EHR system of the tertiary care hospital in Singapore comprised de-identified information on demographics, laboratory tests, drugs, visits, vital signs, diagnoses, and mortality of 5,199 Type 2 DM patients that spanned between January 2011 and February 2018. It should be highlighted here that although our raw data had information on surgical procedures, provider details, patient location, and devices used, etc., we did not extract and standardize them to OMOP CDM standard in this phase of data standardization exercise. This is because our research requirements did not necessitate the use of such information. However, to generate deeper insights regarding the subjects, we extracted the data captured through the cognitive and depression assessment questionnaires. They include the Repeatable Battery for Assessment of Neuropsychological Status (RBANS), the Geriatric Depression Scale (GDS), and the Mini-Mental State Exam (MMSE). Data for all

Table 1 Dataset overview

Item	Details
Dataset name	SG_T2DM
Dataset description	Clinical information of T2DM patients extracted from the EHR system of a tertiary care hospital in Singapore
Data duration	2011–2018
No. of patients	5,199
Median age	58
Gender (male)	54.90%
Available data domains	Conditions, laboratories, drugs, demographics, cognitive, and depression assessment data
Available visit types	Inpatient, outpatient, and emergency

Abbreviations: EHR, electronic health records; T2DM, type 2 diabetes mellitus.

types of visits such as inpatient, outpatient and emergency between 2011 and 2018 were extracted and provided to us in the form of CSV files that contained date and time-stamped clinical information. ►**Table 1** provides the characteristics of our dataset.

This study is approved by the institutional review board (study reference number: 2017/00662).

Observational Medical Outcomes Partnership Common Data Model

OMOP CDM is a patient centric model developed by the OHDSI community that allows to store data about patients across different domains.[7] It has more than 35 tables structured under different domains such as clinical, health system, health economics, metadata, vocabulary, and derived elements. More details on the list of tables under each domain, CDM conventions for populating data under those tables can be found in the online OMOP CDM documentation.[17]

We implemented OMOP CDM version 5.3.1 and populated data only for eight clinical tables such as person, measurement, drug_exposure, visit_occurence, condition_occurrence, observation, observation_period, and death due to our source data availability. In addition, we used standardized algorithms and vocabulary files provided by the OHDSI community to populate the derived tables (drug_era, condition_era) and vocabulary tables. Rest of the OMOP CDM tables are not populated due to a lack of source data.

Transformation of SG_T2DM to Observational Medical Outcomes Partnership Common Data Model

Transformation of our SG_T2DM dataset to OMOP CDM standard involved two steps:

- Content standardization
- Data structure standardization

Content Standardization

Each country has its own medical vocabularies that could only be relevant in their region. For instance, the generic drug glibenclamide is available as Daonil in Singapore, as Diabeta or Glycron in the United States, and as Euglucon in Canada.[18,19] The difference in terminologies between institutes prevents data interoperability, hindering network-based research. OMOP CDM, through its standardized vocabularies, can harmonize data contents from different institutes and unlocks global level health care analytics. OHDSI's Athena is an online resource for OMOP CDM standardized vocabularies. It contains terms from more than 70 vocabularies with a complete mapping of them to the standard concepts.[20] For example, OHDSI Athena provides a vocabulary called "RxNorm Extension" that accommodates the variations in drug names such as brand names, package sizes, drug forms, and manufacturers or distributors, etc. from different countries.

We used Usagi,[21] an open-source semi-automated tool, to map our source terms in the English language to OMOP CDM standard. It suggests relevant standard concepts for our local terms based on textual similarity and assigns a score. Similarity score of 1 indicates an exact match while similarity score of 0 indicates no match.

We chose 0.6 as a threshold for manual review based on prior literature,[10] which used a comparable textual similarity approach. While the prior literature used a score of 400 on a scale of 1,000, instead of choosing an equivalent 0.4 on a scale of 1, we chose 0.6 as threshold based on our detailed scan of the Usagi output and to ensure that majority of our source terms be manually reviewed by the domain expert. Therefore, we categorized our mapping results into two levels as shown below:

- Level 1: raw terms with a similarity score $\geq$0.6
- Level 2: raw terms with a similarity score $<$0.6

While level 2 terms are completely reviewed by our domain expert with pharmacy background, level 1 terms are validated by the informaticians for textual similarity and any mapping discrepancies identified in level 1 are forwarded to the domain expert for further review. The ►**Supplementary Table S1** (available in the online version) shows sample of such mapping discrepancies identified under level 1. Although we hired an expert with pharmacy background to review our terms, institutes willing to adopt OMOP CDM can also consider people with Bachelor of Medicine, Bachelor of Surgery (MBBS), or Doctor of Medicine (MD) background to review their source terms.

Additionally, in scenarios where our source term has multiple standard concept mappings in Athena, we use such combination of concepts to retain the accurate meaning of our source term. For instance, the term "unspecified complication of pregnancy" had two related mappings in Athena which are "finding related to pregnancy" and "complication occurring during pregnancy." Therefore, we use both these concepts to represent our source terms.

Data Structure Standardization

Transforming the raw source data to OMOP CDM standard also involved mapping source data attributes to the correct columns in the appropriate OMOP CDM tables. OMOP CDM schema, comprising of more than 35 tables, was created by using the Data Definition Language statements available in the OHDSI GitHub repository[17] and made available as an indexed database.

Before mapping our raw data fields to OMOP CDM tables to identify and drop records of low quality from source data, we applied a set of exclusion rules: (1) missing or useless values in patient identifiers; (2) multiple gender values for a patient; (3) multiple date of birth records for a patient with a difference of more than 2 years; (4) date of birth <1900; (5) event start_date is greater than event end_date; (5) records with only patient identifier but missing other relevant variables such as event_dates and values for that specific event; and (6) removal of duplicates and useless values. Useless values are defined as data entry errors that are present in the source data. For instance, our source data had terms such as "ADM OVERRIDE," "ADM OVERRIDE OVERRIDE," "LIS OBR24," etc. with no other accompanying information, rendering them useless for the purpose of analysis.

In practice, we know that poor quality records can lead to poor decisions. Therefore, we eliminated such records from our source data to improve decision-making and minimize the burden on researchers to handle data quality issues. Additionally, through our Extract Transformation and Load (ETL) process, we handled inconsistencies such as incorrect date formats, missing values for mandatory columns, adaptation of data types, separation/combination of attribute values, creation of new derived attributes, expansion of source terms abbreviation, etc. to produce consistent data as per OMOP CDM specifications. The source data were then transformed to OMOP CDM standard as per the CDM specifications listed in OHDSI github[17] by using Python, R and Structured Query Language (SQL) codes. ►Fig. 1 illustrates the data field mapping between a raw patient file and person CDM table. Fields such as "religion," "Marital_Status," and "language" are not mapped because there was no matching column in the person table. However, we followed OMOP CDM convention to record these information in the "observation" table.

For questionnaire data, every question in the cognitive and depression assessment questionnaires was treated as a distinct observation source value and every question along with the patient response was treated as a row in the CDM observation table. Unlike EHR data, our questionnaire data had local terms that were not found in OHDSI Athena. Therefore, we followed the OHDSI recommended procedure, created custom concepts to represent our local questionnaire terms, and assigned them a concept ID greater than 2 billion to avoid conflict with any of the existing OMOP CDM concepts. We created 123 custom concepts to represent 89% of our questionnaire terms that did not have any matching OMOP concept, and they are added to our local vocabulary tables such as concept, concept_relationship, concept_ancestor and source_to_concept_map. The ►Supplementary

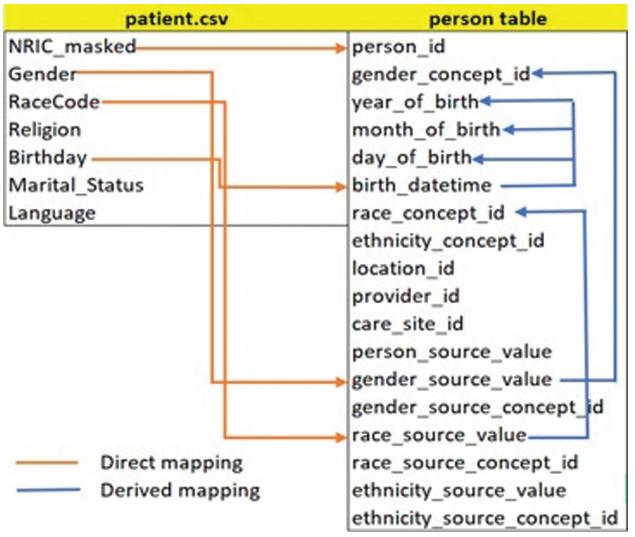

Fig. 1 Data field mapping between raw patient data and Observational Medical Outcomes Partnership Common Data Model person table.

Table 2 Raw cognitive and depressive assessment data from the source csv file

Patient_ID	List recall item 1 market	List recall item 2 apple
123	1	0
456	0	1

Table 3 After pivoting the raw survey data structure for easy interpretability and manipulation

Patient_ID	Question	Response	Response_string
123	List recall item 1 market	1	Yes
456	List recall item 2 apple	0	No

Table S2 (available in the online version) lists the step-by-step procedure to create custom concepts.

►Table 2 shows the format of our raw questionnaire data. The column headers indicate the question, and cell value indicates the patients' response to the question. For example, as a part of cognitive assessment, patient was asked to listen and repeat a list of words that was read by the examiner. For instance, if an interviewer read out 10 words to the patient and only 5 words are recalled (repeated) successfully by the patient, then patient gets a total score of 5 out of 10. Each successful recollection of a word from the list yields a score of 1 irrespective of the sequence in which words were recalled. ►Table 3 shows the pivoted form of raw questionnaire data. We pivot the raw questionnaire data structure for ease of interpretation and manipulation. Finally, ►Table 4 shows the questionnaire data standardized as per the CDM convention. For example, Observation_concept_ID such as "2000000368" indicate the custom concept ID created by us to indicate the question, which is "list recall item 1 market."

Table 4 Cognitive assessment data transformed as per Common Data Model convention of observation table using custom concepts

Person_ID	Observation_concept_ID	Value_as_number	Value_as_string	Value_as_concept_ID	Observation_source_value
123	2000000368	1	Yes	4188539	List recall item 1 market
456	2000000369	0	No	4188540	List recall item 2 apple

Table 5 Source data exclusion rules

	Source data table (source count)						
Exclusion rule	Person (5,214)	Condition (338,688)	Drug (2,274,749)	Visit (261,499)	Laboratory (11,563,678)	Death (244)	Observation (542,211)
Missing or useless values in patient identifiers	0	0	0	0	0	0	0
Multiple gender values for a patient	0	NA	NA	NA	NA	NA	NA
Multiple date of birth records for a patient with a difference of more than 2 y	0	NA	NA	NA	NA	NA	NA
Date of birth < 1900	0	NA	NA	NA	NA	NA	NA
Event start_date > event end_date (ex: Drug_start_date > Drug_end_date etc.)	0	0	0	0	0	0	0
Missing of multiple relevant variables	0	44,453	0	840	522,455	0	324,820
Removal of duplicates and junk values	15	7,020	62	9,304	612,465	0	0
Final cleaned source data count	5,199	331,669	2,274,687	251,355	10,428,758	244	217,391

Abbreviation: NA, not applicable.

Similarly, value_as_concept_ID represents the existing OMOP concept IDs for terms "yes" and "no." It should be highlighted here that our raw questionnaire terms did not violate the CDM constraints on length of field.

The final dataset after transformation was uploaded to OMOP CDM database by adhering to the order of populating tables as defined by OMOP CDM constraints.

Evaluation of Transformation

We validated our transformation using three approaches: (1) assessing mapping coverage, computing the percentage of source terms that were able to be represented as concepts in OMOP CDM form; (2) local raw dataset versus CDM dataset analysis, comparing the key demographic and clinical characteristics between local raw data and CDM data using descriptive statistics; and (3) use of Data Quality Dashboard[15] to run exhaustive validation checks on the conformance, completeness, and plausibility of data in the CDM dataset.

Furthermore, we validated our implementation of OMOP CDM instance and its accompanying tool for analysis called Atlas[22] is fit for use in research by executing a simple treatment pathways study, which was inspired by an already published study by George Hripcsak et al.[23] Treatment pathways refer to the series of interventions a person received for a period of a time. It should be highlighted that this study is conducted only as a proof of concept to demonstrate the usefulness of OMOP CDM and not to generate any medically relevant findings.

Results

Mapping SG_T2DM to Observational Medical Outcomes Partnership Common Data Model

The SG_T2DM cohort used in this study is restricted only to one hospital. While the raw data had 5,214 patients, only 5,199 patients were transformed to OMOP CDM standard. The decrease in patient count was due to the exclusion of duplicate patient records from the source data. Similarly, we applied further exclusion rules to other tables to identify and drop records of low quality. ►**Table 5** shows the effect of deletions of such records from source tables.

These 5,199 patients contributed 13,504,104 rows of clinical information on diagnosis, laboratory tests, visits, drugs, observation, and death. In terms of content standardization, we specifically mapped 4,806 unique diagnosis terms, 1,600 unique laboratory terms, 2,592 unique drug terms, and 149 unique questionnaire terms to OMOP CDM standard. ►**Fig. 2** shows the number of terms under each level for different domains of data. Examples of terms under each level for different domains can be found under ►**Supplementary Table S3** (available in the online version).

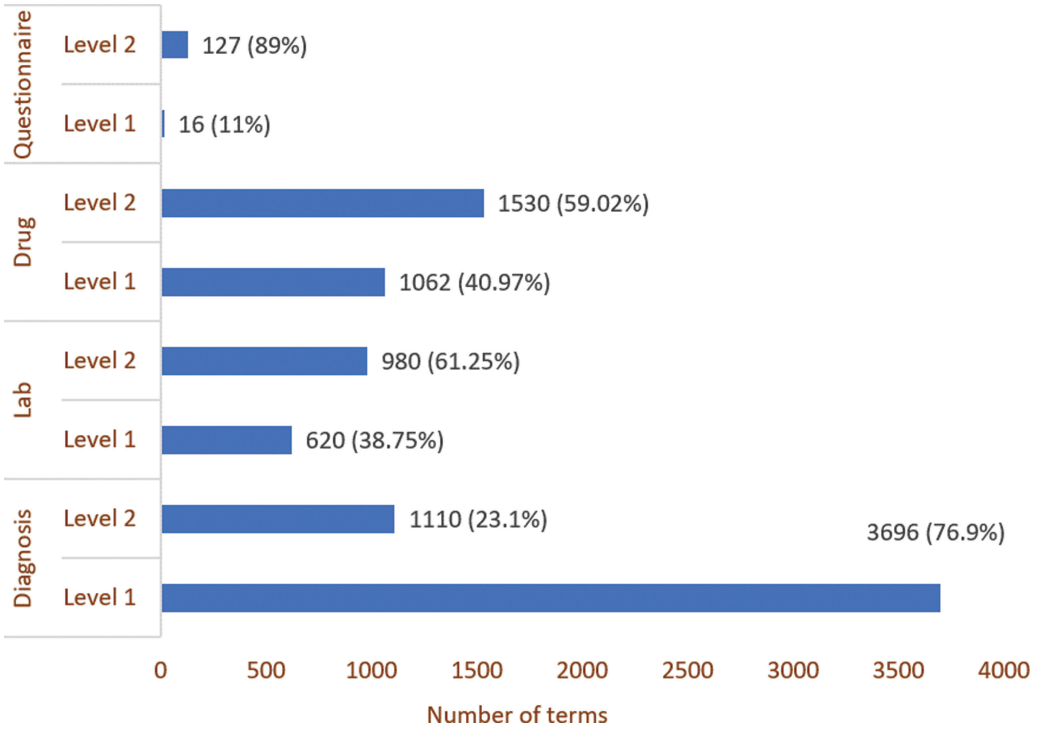

Fig. 2 Number of terms under each level for different data domains. While diagnosis domain had majority of their terms under level 1, laboratory, drugs, and questionnaire data had majority of their terms under level 2.

Table 6 Mapping coverage of clinical and questionnaire data

Domain	No of unique terms	Number of unique unmapped terms	Total number of events	Unmapped term events as % of total events (data)
Condition	4,806	24	331,699	4.54
Labs & vital signs	1,617	121	10,428,758	4.49
Drugs	2,592	12	2,274,687	0
Questionnaire	143	123	217,391	86

Additionally, our DE was not able to map all source EHR terms to OMOP CDM concepts. However, unmapped terms were minimal as shown in ►Table 6. Instead of dropping the unmapped records, we leveraged the "_source_value" column of OMOP CDM tables to store the source terms and assigned a concept ID of 0. This enables the researchers to conduct analysis based on local source terms if required. Unlike EHR data, questionnaire data, specifically RBANS and MMSE had lot of unmapped terms as shown in ►Fig. 3. It should be highlighted here that unmapped terms indicate the terms for which we did not find any logically relevant mapping in OHDSI Athena. The ►Supplementary Tables S4 (available in the online version) shows the sample of our source terms from different domains that were mapped and not mapped to OMOP CDM standard.

Local Raw versus Common Data Model Dataset Comparison

In line with the previous research on assessing feasibility of OMOP CDM,[8] we also extracted and compared the information on key demographics, clinical, and lifestyle factors between raw source data and target CDM data. The results shown in ►Table 7 indicate that the characteristics are consistent before and after transformation.

Implementation of Harmonized Intrinsic Data Quality Framework

We used "Data Quality Dashboard," an open-source R package developed based on HIDQF by the OHDSI community to assess the quality of our transformed dataset.[15] We executed more than 2,500 validation checks based on conformance, completeness, and plausibility of our data in OMOP CDM. ►Fig. 4 shows the results from our data quality Dashboard. The list of checks used by Data Quality Dashboard can be found from the github repository of Data Quality Dashboard.[24]

It helped us uncover 38 issues out of 2,622 validation checks evaluated in two different ways such as verification and validation across different dimensions such as conformance, completeness, and plausibility. Definitions of these

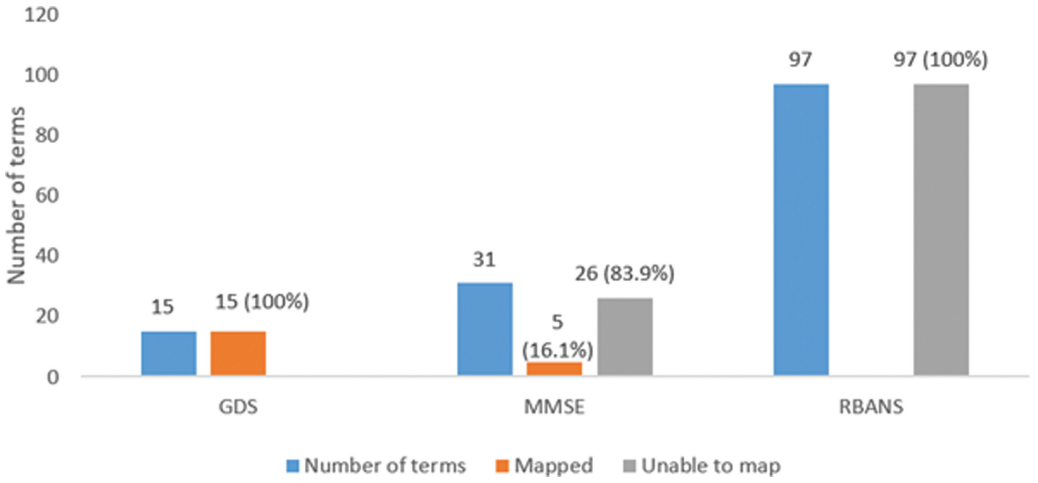

Fig. 3 Mapping coverage of questionnaire data. Repeatable Battery for Assessment of Neuropsychological Status and Mini-Mental State Exam had minimal coverage when compared with Geriatric Depression Scale, which has 100% matching terms.

Table 7 Comparison of characteristics between raw data and Observational Medical Outcomes Partnership Common Data Model data

	SG_T2DM raw data	SG_T2DM OMOP CDM data
n	5,199	5,199
Demographics		
Female (%)	45.08	45.08
Chinese (%)	51.7	51.7
Lifestyle		
Ex-smoker (%)	8	8
Current smoker (%)	7	7
Nonsmoker (%)	54.74	54.74
BMI category (kg/m^2) (%)		
<18.5	0.8	0.8
18.5–23.0	11.5	11.5
>23.0–27.5	26.73	26.73
>27.5	30.77	30.77
Clinical measures mean (SD)		
BMI (kg/m^2)	27.54 (5.21)	27.54 (5.21)
SBP (mmHg)	134.35 (26.58)	134.35 (26.58)
DBP (mmHg)	68.95 (15.52)	68.95 (15.52)
FiO_2 (%)	46.46 (22.20)	46.46 (22.20)
eGFR (mL/min/1.73 m^2)	34.55 (36.12)	34.55 (36.12)
Triglycerides (mmol/L)	1.94 (1.78)	1.94 (1.78)
HDL-C (mmol/L)	1.23 (0.38)	1.23 (0.38)
LDL-C (mmol/L)	2.80 (0.98)	2.80 (0.98)
HbA1c (%)	8.3 (1.92)	8.3 (1.92)

Abbreviations: BMI, body mass index; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; FiO_2, fraction of inspired oxygen; HDL-C, High-density lipoprotein cholesterol; LDL, low-density lipoprotein cholesterol; SBP, systolic blood pressure.

terms can be found online from the BookOfOHDSI.[25] A total of 47% of our errors are due to some of our laboratory tests violating their plausible value limit; 13% of errors are due to invalid visit_ID, which indicates that patients have visit_IDs that are not present in the master visit table; and 13% of errors are due to lack of standard concept ID (assigned 0) across columns such as "unit_concept_ID" and "route_concept_ID." A total of 13% of errors are due to the use of concept_IDs in different clinical tables that do not adhere to their respective domain conventions; 7.8% of errors are due to lack of source codes from our data in the concept table; and 2.63% of errors are due to missing observation periods.

Observational Medical Outcomes Partnership Common Data Model Instance: Fit for Use Assessment

We validated our OMOP CDM setup, and its analysis tool called Atlas[22] for observational research by running a simple treatment pathways study, inspired by the study published by George Hripcsak et al.[23] The objective of executing this study was only to assess the utility of OMOP CDM database and atlas tool to design cohort definitions, create cohorts of interest, and conduct analysis using Graphical User Interface (GUI) options available, etc. and not to generate any medically relevant findings. Therefore, discussion on clinical relevance of our results is considered out of scope of this paper and will be pursued as a separate publication in future. This section will report differences of our study cohort criteria from the original study, treatment sequence of our patients, and time taken to run their analysis.

First, our study is different from the original study on multiple factors: (1) the original study generated treatment pathways for multiple conditions such as hypertension, depression and diabetes, but we generated treatment sequence only for Diabetes patients due to our data availability. (2) The original study had a criteria of continuous drug exposure of every 120 days to be qualified for respective event cohorts, whereas we chose occurrence of drug records anytime in patient's records after index date to be qualified

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1910	18	1928	99%	205	0	205	100%	2115	18	2133	99%
Conformance	257	15	272	94%	47	0	47	100%	304	15	319	95%
Completeness	158	4	162	98%	7	1	8	88%	165	5	170	97%
Total	2325	37	2362	98%	259	1	260	100%	2584	38	2622	**99%**

Fig. 4 Dashboard presenting our data quality results. A total of 1.4% out of 2,622 checks have failed in our dataset.

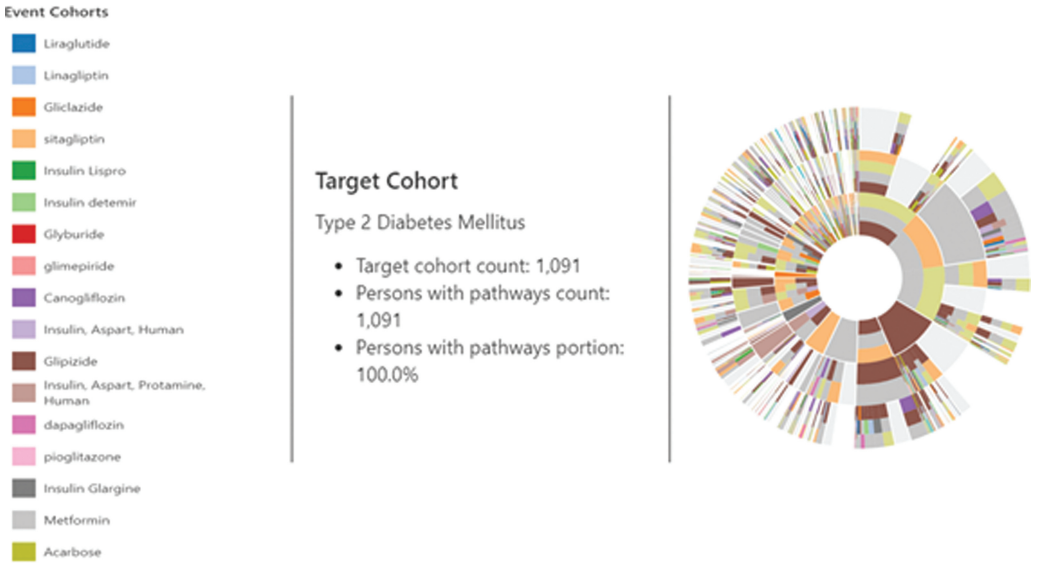

Fig. 5 Sunburst plot indicates the treatment pathways for 1091 Type 2 diabetes mellitus patients in our dataset. First circle represents first-line of medication and second circle indicates second line of medication and so on and so forth.

for the event cohort. (3) The original study used 23 different items under Type 2 diabetes mellitus (T2DM) medication classes, whereas we used only 7 oral antidiabetic drug classes along with Insulin variants due to our data characteristics. The ►**Supplementary Table S13** (available in the online version) contains treatment pathways cohort definitions used for our research. Additionally, we configured five days as a combination window, indicating the time interval when two event cohorts need to overlap to be considered a combination. Moreover, we also consider an event to be counted in the pathway only if there are at least five subjects in the target cohort who meets the criteria of event cohort.

►**Fig. 5** shows the sunburst plot indicating the treatment sequences of 1,091 patients who satisfied our cohort entry criteria of having 1 year of observation period and 3 years of following period. Through generated results, we can infer based on the color codes that majority of our patients had metformin either in combination with other diabetic drugs or as the only first-line medication which aligns with the first-line recommended treatment by the American Association of Clinical Endocrinologists diabetes treatment algorithm.[26] To run this analysis, we used Atlas version 2.7.3 to create a target cohort of T2DM patients and event cohorts for different diabetic medications. The execution of this analysis—after target and event cohort creation—took 0.31 seconds, which is found from the Atlas executions tab of Cohort pathways menu. Nonetheless, study activities such as import of JSON file from an external source, creation, addition of cohorts with concept sets using Atlas features, and successful generation of results indicate that our OMOP CDM instance is fit for the purpose of local and network research using OMOP CDM data.

Discussion

In this study, we standardized EHR, cognitive, and depression assessment questionnaires data from a tertiary care hospital in Singapore. The results indicate the usefulness of OMOP CDM for standardization of such disparate data sources. In this section, we will discuss our experience including challenges and benefits of adopting OMOP CDM for conversion of SG_T2DM dataset from Singapore.

First, we found out that the OMOP CDM vocabularies is very exhaustive with respect to clinical data as majority of our source terms from EHR data are mapped to OMOP concepts as highlighted earlier. Detailed investigation of unmapped terms from raw diagnosis and laboratories data revealed that most of them are administrative artifacts and form only a small portion of our data. Top five unmapped

terms from each domain can be found in ►Supplementary Table S5 (available in the online version). For diagnosis data, the administrative artifacts such as "unspecified reason for consultation" or "Medical care, unspecified," etc. can be stored under observation table as per CDM convention. Similarly, for drugs data, we found that terms which are unmapped are due to the health care supplements records in our data. It should be highlighted here that CDM convention states that nutritional products/supplements should be stored under device domain instead of drug domain due to the questionable effects of supplements on the body. Despite our efforts to map health supplement terms to concepts under device domain, not all of them were able to be mapped to OMOP CDM standard. Nonetheless, such terms are very few as shown in ►Supplementary Tables S6–S10 (available in the online version) and did not affect our analysis. A possible approach to address this challenge would be to add our local supplements terms to OHDSI vocabularies. However, this is not attempted at this stage considering the negligible impact of our unmapped drug terms. On the other hand, for cognitive and depression assessment questionnaire data, we were able to find exact matching OMOP concepts only for terms from GDS questionnaire. The content coverage of OMOP CDM vocabularies for RBANS and MMSE questionnaire is very minimal as shown earlier. The ►Supplementary Tables S11 and S12 (available in the online version) shows the sample of terms from RBANS and MMSE questionnaires that cannot be mapped to OMOP CDM. Though we find RBANS subtest headers such as memory recall, digit span, etc. available in OMOP CDM vocabularies, the questions under the subtests useful for our research are not supported by OMOP CDM vocabularies as of now. Therefore, we adopted OHDSI's recommended procedure to represent local terms (in this case questionnaire terms) by creating custom concepts. However, to make our questionnaire terms publicly available as standard concepts in OHDSI's Athena, we have become a member of the psychiatry working group in the OHDSI community to understand the possibilities of standardizing such questionnaire terms to benefit the community. Moreover, these questionnaire terms are found publicly available online and we do not foresee any license constraints in adding them to OMOP CDM vocabulary.

Second, the data structure standardization involved significant time investment upfront to understand the OMOP CDM conventions, source data and its attributes to map them to the appropriate columns in OMOP CDM tables. This task was performed in collaboration with our data owners who resolved our queries with respect to the source data flow and its attributes. Additionally, the ETL programming codes and vocabulary mapping files created by the DE will be reused for our future data loads and will not require considerable resources as it took during initial setup. The extra effort on ETL and vocabulary mapping will only be for the brand-new raw data tables and source terms that might be part of the future data loads. We also realized that by standardizing our data to OMOP CDM standard, our understanding of the source data increased multifold. This helped us identify and resolve data quality issues upfront, thereby preventing major data quality issues from the source data being transferred to the OMOP CDM data.

Third, comparison of key demographic, lifestyle, and clinical factors between local raw dataset and CDM based dataset resulted in consistent findings. In addition, use of DQD offered rich insights on the transformation of our dataset. The errors that are due to violations of plausible laboratory value limit is mainly because the threshold value limits are not yet customized to the Singapore context due to the significant resources required to assemble a group of experts to arrive at a threshold values for different laboratory measurements. Hence, plausibility checks for measurements are currently based on existing expert driven limits configured in DQD. The invalid visit_IDs are mainly because our raw data had some discrepancies. For instance, none of our questionnaire data which is stored in observation table had any visit_IDs and was assigned 0. Similarly, 15.75% of diagnosis data, 8.96% of laboratory data, and 6.13% of our drugs data had a visit_ID, which was not present in the master visit table of patients' records. Errors due to lack of standard concept_ID for "unit_concept_ID" column is expected because our observation table had questionnaire data and the corresponding "unit_concept_ID" column in the observation table was filled with zeroes because it was not applicable for questionnaire data. Additionally, 6.04% of our condition source codes and 1.24% of our route of administration source codes were not present in the concept table. One possible solution to resolve this issue could be to update the vocabulary version in our OMOP CDM instance. Finally, patients with missing observation periods were due to lack of any clinical information regarding those patients. For instance, we only had their demographic information but did not have any clinical information to determine the observation period for them. Upon investigation of all these 38 errors through DQD and by querying the database using SQL, it was clear to us that their impact on analysis is negligible and as such are not addressed at this stage.

Furthermore, one of the main benefits of adopting OMOP CDM is to facilitate large-scale observational research. To validate the usefulness of our OMOP CDM instance for participation in research-based activities, we used Atlas application to import publicly available json files, search for concepts, and define cohorts required for our treatment pathways study. We were able to successfully generate treatment pathways for diabetes disease. Atlas allowed us to experiment with different analysis settings such as "combination window" and "minimum cell count," helping us to better understand the variations in treatment pathways through its easy-to-use interface. Additionally, being OMOP CDM compliant enabled us to participate in network studies across the globe.[27,28] During the network studies, the analysis code written at the coordinating center was reused at our site to generate results without sharing of data. In other words, it brought analysis codes to the data. Thus, being able to test, the same research question across multiple sites with different data characteristics increases the external validity of research.

Despite the benefits of adopting OMOP CDM, one major challenge that we encountered during the transformation of raw data to OMOP CDM is the required resources to map raw source terms to OMOP CDM terms. The level 1 source terms required little to no manual review as there were equivalent terms in the CDM vocabularies. However, most of our source terms were under level 2 and required expert review. It should be highlighted here that this is a mandatory and labor-intensive process that any healthcare institute adopting OMOP CDM must go through.

Limitation

Our study is not free from limitations. First, our dataset size is small, only 5,199 patients. Second, our standardized dataset does not contain information on procedures, devices, location, and unstructured data. Therefore, we did not assess the feasibility of OMOP CDM for representing such data. However, we hope to include unstructured data and other clinical information discussed above in our future data loads.

Conclusion

In this study, we demonstrate that it is feasible to transform EHR data from a tertiary care clinic in Singapore to OMOP CDM standard. Though there was a minimal information loss, data were found to be of sufficient quality for our research requirements. However, challenges to standardize cognitive and depression assessment questionnaire data to OMOP CDM exist and requires further work. Nonetheless, the standardization of SG_T2DM provided us a deeper understanding of the source data and enabled us to participate in large-scale federated observational research using EHR data. Considering the increased interest among other health care institutions in Singapore to adopt OMOP CDM as a platform for clinical research, we believe that the results of this study can guide other institutes in their OMOP CDM journey.

Clinical Relevance Statement

Adoption of OMOP CDM can facilitate network-based research without any hassle. In addition, standardizing our dataset with such a wide range of information can support Singapore Ministry of Health's goal by offering valuable insights to the broader research community on the epidemiology of the T2DM; whether certain treatments have better outcomes; impact of lifestyle on development of disease over time; and how it affects mental health.

Multiple Choice Questions

1. Which of the following tool is provided by the OMOP/OHDSI community for vocabulary mapping?
   a. Usagi
   b. White rabbit
   c. Rabbit in a hat
   d. Achilles

   Correct Answer: The correct answer is option a. Rest of the options indicate tools used for database characterization (Achilles), analyze the structure and contents of a database as preparation for designing ETL (White Rabbit), and interactive tool for designing ETL (rabbit-in-a-hat). Usagi is the semi-automated tool widely used by the community for vocabulary mapping.

2. Which of the following CDM table is used to store patient's social and lifestyle information?
   a. Condition_occurrence
   b. Observation
   c. Person
   d. Drug_exposure

   Correct Answer: The correct answer is option b. As per CDM convention, observation table is used to store clinical facts about a person obtained in the context of examination, questioning or a procedure. Any data that cannot be represented by any other domains, such as social and lifestyle facts, medical history, family history, etc. are recorded here.

Protection of Human and Animal Subjects

We used de-identified patient data for this study and is approved by the Institutional Review Board (Study Reference Number: 2017/00662).

Note

E.S.T. and S.C.L. are co-investigator on grants from the NMRC under the OF-LCG and CG schemes. The grants are awarded to the institution which employ them. S.M.K.S. current research team member was hired under this grant. J.Y.S. former research team member was hired under this grant. S.Y.M. current research team member is working in the institution which received the grant from the NMRC under the OF-LCG and CG schemes.

Funding

This research is funded by the National Medical Research Council (NMRC) under the Open Fund - Large Collaborative Grant (OF-LCG) - NMRC/OFLCG/001/2017 and Centre Grant (CG) schemes - NMRC/CG/C016/2017.

Conflict of Interest

None declared.

Acknowledgments

The authors would like to thank Joel Sim, an experienced pharmacist, who performed the vocabulary mapping task. In addition, they are grateful to Dr. Keven Ang, Dr. Wang Jiexun, and Ms. Clara Chan for their valuable inputs during data extraction which helped us understand the source data better.

References

1 Al-Lawati JA. Diabetes mellitus: a local and global public health emergency!. Oman Med J 2017;32(03):177–179
2 Diabetes. Who.int. Accessed April 6, 2021 at: https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=WHO%

20estimates%20that%20diabetes%20was,onset%20of%20type%202%20diabetes
3 MOH | News Highlights. Moh.gov.sg. Accessed April 6, 2021 at: https://www.moh.gov.sg/news-highlights/details/diabetes-the-war-continues
4 Luo M, Tan LWL, Sim X, et al. Cohort profile: the Singapore diabetic cohort study. BMJ Open 2020;10(05):e036443
5 Munshi MN. Cognitive dysfunction in older adults with diabetes: what a clinician needs to know. Diabetes Care 2017;40(04):461–467
6 Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333–341
7 Data Standardization – OHDSI. Ohdsi.org. Accessed April 6, 2021 at: https://ohdsi.org/data-standardization/
8 Haberson A, Rinner C, Schöberl A, Gall W. Feasibility of mapping austrian health claims data to the OMOP common data model. J Med Syst 2019;43(10):314
9 Maier C, Lang L, Storf H, et al. Towards implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform 2018;9(01):54–61
10 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(01):1110
11 Lamer A, Depas N, Doutreligne M, et al. Transforming French Electronic Health Records into the observational medical outcome partnership's common data model: a feasibility study. Appl Clin Inform 2020;11(01):13–22
12 Zhou X, Murugesan S, Bhullar H, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf 2013;36(02):119–134
13 Cho S, Sin M, Tsapepas D, et al. Content coverage evaluation of the OMOP vocabulary on the transplant domain focusing on concepts relevant for kidney transplant outcomes analysis. Appl Clin Inform 2020;11(04):650–658
14 Patrick Ryan, Martijn Schuemie, Vojtech Huser, Chris Knoll, Ajit Londhe and Taha Abdul-Basser (2019). Achilles: Creates Descriptive Statistics Summary for an Entire OMOP CDM Instance. R package version. Accessed 2019 at: https://ohdsi.github.io/Achilles/1.6.7
15 OHDSI/DataQualityDashboard. GitHub. Published 2021. Accessed April 6, 2021 at: https://github.com/OHDSI/DataQualityDashboard
16 Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(01):1244
17 OHDSI/CommonDataModel. GitHub. Published 2021. Accessed April 6, 2021 at: https://github.com/OHDSI/CommonDataModel/
18 Glyburide DrugBank Online. Go.drugbank.com. Published 2021. Accessed April 6, 2021 at: https://go.drugbank.com/drugs/DB01016
19 Recommendations to avoid use of glibenclamide in the elderly and renal-impaired. HSA. Published 2021. Accessed April 6, 2021 at: https://www.hsa.gov.sg/announcements/safety-alert/recommendations-to-avoid-use-of-glibenclamide-in-the-elderly-and-renal-impaired
20 Athena. Athena.ohdsi.org. Published 2021. Accessed April 6, 2021 at: https://athena.ohdsi.org/search-terms/start
21 OHDSI/Usagi. GitHub. Published 2021. Accessed April 6, 2021 at: https://github.com/OHDSI/Usagi
22 ATLAS. Atlas.ohdsi.org. Published 2021. Accessed April 6, 2021 at: https://atlas.ohdsi.org/
23 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
24 OHDSI/DataQualityDashboard. GitHub. Accessed April 6, 2021 at: https://github.-com/OHDSI/DataQualityDashboard/tree/master/inst/csv
25 Observational Health Data Sciences. Informatics. The book of OHDSI. Github.io. Accessed April 6, 2021 at: https://ohdsi.github.io/TheBookOfOhdsi/
26 AACE/ACE Clinical Practice Guidelines for Developing a Diabetes Mellitus Comprehensive Care Plan - © 2015. Accessed April 6, 2021 at: pro.aace.com. https://pro.aace.com/disease-state-resources/diabetes/clinical-practice-guidelines/aaceace-clinical-practice-guidelines
27 Huser V. Data quality assessment of laboratory data. Accessed 2021 at: https://knowledge.amia.org/72332-amia-1.4602255/t005-1.4604904/t005-1.4604905/3413748-1.4605506/3413748-1.4605507?qr=1
28 Jonnagaddala J. External validation of type II diabetes electronic phenotyping algorithms. Accessed 2021 at: Presented at the: https://www.ohdsi.org/wp-content/uploads/2020/05/OHDI_-symposium_2020_T2DM_poster_JJ.pdf
29 Report S. Data protection in the internet. In: Data Protection in the Internet. SpringerCham2019https://www.springer.com/gp/book/9783030280482. Accessed January 18, 2021
30 Rodrigues JJ, de la Torre I, Fernández G, López-Coronado M. Analysis of the security and privacy requirements of cloud-based electronic health records systems. J Med Internet Res 2013;15(08):e186
31 Salloway MK, Deng X, Ning Y, et al. A de-identification tool for users in medical operations and public health. 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). Accessed 2021 at: Doi: 10.1109/bhi.2016.7455951
32 Hripcsak G, Mirhaji P, Low AF, Malin BA. Preserving temporal relations in clinical data while maintaining privacy. J Am Med Inform Assoc 2016;23(06):1040–1045
33 Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3(01):160035