Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study

Antoine Lamer[1] Nicolas Depas[1] Matthieu Doutreligne[2] Adrien Parrot[3,4] David Verloop[5] Marguerite-Marie Defebvre[5] Grégoire Ficheur[1] Emmanuel Chazard[1] Jean-Baptiste Beuscart[1]

[1] Univ. Lille, CHU Lille, ULR 2694-METRICS: Évaluation des Technologies de Santé et des Pratiques Médicales, F-59000 Lille, France
[2] Bureau Etat de Santé de la Population, Ministère des Affaires Sociales et de la Santé, Direction de la Recherche, des Etudes et des Statistiques - Observation de la Santé et de l'Assurance Maladie, Paris, France
[3] Université Paris Descartes, Paris, France
[4] Web INnovation Données-Direction des Systèmes d'Information, Assistance Publique - Hôpitaux de Paris, Paris, France
[5] Service Etudes et Statistiques, ARS Hauts-de-France, Lille, France

Address for correspondence Antoine Lamer, PhD, CIC-IT 1403 de Lille, Institut Coeur Poumon, Bd du Professeur jules Leclercq, 3ième étage Aile Est, CS 70001, Lille Cedex 59037, France (e-mail: antoine.lamer@chru-lille.fr).

Appl Clin Inform 2020;11:13–22.

Abstract

Background Common data models (CDMs) enable data to be standardized, and facilitate data exchange, sharing, and storage, particularly when the data have been collected via distinct, heterogeneous systems. Moreover, CDMs provide tools for data quality assessment, integration into models, visualization, and analysis. The observational medical outcome partnership (OMOP) provides a CDM for organizing and standardizing databases. Common data models not only facilitate data integration but also (and especially for the OMOP model) extends the range of available statistical analyses.

Objective This study aimed to evaluate the feasibility of implementing French national electronic health records in the OMOP CDM.

Methods The OMOP's specifications were used to audit the source data, specify the transformation into the OMOP CDM, implement an extract–transform–load process to feed data from the French health care system into the OMOP CDM, and evaluate the final database.

Results Seventeen vocabularies corresponding to the French context were added to the OMOP CDM's concepts. Three French terminologies were automatically mapped to standardized vocabularies. We loaded nine tables from the OMOP CDM's "standardized clinical data" section, and three tables from the "standardized health system data" section. Outpatient and inpatient data from 38,730 individuals were integrated. The median (interquartile range) number of outpatient and inpatient stays per patient was 160 (19–364).

Conclusion Our results demonstrated that data from the French national health care system can be integrated into the OMOP CDM. One of the main challenges was the use of international OMOP concepts to annotate data recorded in a French context. The use of local terminologies was an obstacle to conceptual mapping; with the exception of an adaptation of the International Classification of Diseases 10th Revision, the French health care system does not use international terminologies. It would be interesting to extend our present findings to the 65 million people registered in the French health care system.

Keywords
- data integration
- secondary use
- observational medical outcome partnership
- Observational Health Data Sciences and Informatics

received
June 27, 2019
accepted after revision
November 25, 2019

DOI https://doi.org/10.1055/s-0039-3402754.
ISSN 1869-0327.

Background and Significance

Over the last few years, several common data models (CDMs) have been implemented in the health care field.[1–7] These data models are designed to integrate data into a common structure, even when data have been collected through distinct and heterogeneous systems. They enable data exchange, sharing, and storage. Consortia behind the CDMs may also provide a set of tools for assessing data quality, integrate data into models, and perform visualization and analysis.

Garza et al evaluated four CDMs:[8] Sentinel v5.0,[1] PCORNet (National Patient-Centered Clinical Research Network) v3.0,[2] observational medical outcome partnership (OMOP v5.0),[5] and Clinical Data Interchange Standards Consortium Study Data Tabulation Model (CDISC-SDTM) v1.4[9] against 11 criteria for content coverage, integrity, flexibility, ease of querying, standards compatibility, and ease and extent of implementation. Each CDM was populated with 300 records from the MURDOCK (Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis) longitudinal, community-based health study. The records capture self-reported data about conditions, hospital visits, procedures, medications, demographic data, socioeconomic indicators, as well as data from health care facilities. The CDM developed by the OMOP consortium gave the best results, particularly with regard to terminology coverage. Conversely, Xu et al[10] compared the OMOP CDM and other models with regard to the integration of claims data but concluded that (1) all the models studied were very similar and (2) the differences in use had a minor impact. Hripcsak et al reported that more than 200 million patients had already been integrated into the OMOP CDM by early adopters of the project, such as the United States, United Kingdom, the Netherlands, Sweden, Italy, Korea, and Taiwan.[11] Along with data standardization, the OMOP model offers a large number of statistical tools (as R packages) that are dedicated to pharmacoepidemiologic research and enable the implementation of validated models (new user cohort, case control, self-controlled case series, self-controlled cohort, disproportionality analysis, temporal pattern discovery, and longitudinal gamma Poisson's shrinker).[12] Today, the OMOP CDM is maintained by the Observational Health Data Sciences and Informatics (OHDSI) consortium.[5]

Objectives

Several researchers have described the conversion of their data into the OMOP CDM format in various contexts.[13–17] However, to the best of our knowledge, the present study is the first to have attempted to replicate the OMOP's findings and to provide codes and specifications for French longitudinal health care data.

Hence, the main objectives of the present study were to evaluate the implementation of data from the French health care system in the OMOP CDM and to provide documentation for reproducing this process.

Methods

Extraction of Data from French National Databases

The data were extracted from the "Système National des Données de Santé" (SNDS) as part of the French nationwide "Personnes Agées en Risque de Perte d'Autonomie" (PAERPA) project deployed by the French Ministry of Social Affairs and Health from October 2014 to December 2019.[18] This experimental program is being implemented in 16 administrative areas, and focuses on frail adults aged 75 years and over. In this population, the patient's health pathway is coordinated by a family physician and involves at least one other health care professional, most often a nurse and/or a community pharmacist. It was agreed from the outset that the evaluation of the project would be based on French health insurance data, and that no clinical data would be collected from patients. The program's implementation in the Hauts-de-France region included several actions with regard to the risk of adverse drug reactions. We therefore considered that the OMOP model was relevant and would be required for a high-quality analysis.[12]

The centralized French national health care database ("Système National des Données de Santé" [SNDS]) has been operating since 1979, and currently contains health information on more than 65 million living people.[19,20] The SNDS data are made available by the "Institut National des Données de Santé" (INDS). The SNDS is based on the following three data sources: (1) the national health insurance database (Système national d'information inter-régimes de l'Assurance maladie [SNIIRAM], which contains data on ambulatory care data, including all outpatient drug delivery, and medical appointments), (2) the national inpatient stay database ("Program de Médicalisation du Système d'Information" [PMSI]) that contains data on inpatient stays in all public- and private-sector hospitals), and (3) the national death registry containing death certificates ("Centre d'Epidémiologie sur les causes médicales de Décès," CépiDc,). The three databases are linked together at the national level in the SNDS. Each data source uses its own patient identifier, and so the INDS creates a unique anonymous identifier that links each item of information to his/her owner (patient). The SNDS data can be extracted for specific research projects. Each extraction is limited to the relevant data fields (columns) and patients (rows). However, all extractions share the same format. The complete extraction of a single year's data for 65 million patients corresponds to around 150 flat files (total volume: ~2.6 Tb). Given the size and quality of this centralized system, we hypothesized that the national-scale exploitation of these data for pharmacoepidemiologic purposes would be of great value.

The main criteria for data extraction were as follows: residence in the Valenciennois-Quercitain area of northern France, and age 75 years or over on January 1, 2015. Data were extracted for the period between January 1, 2014 and December 31, 2017. The following items of information were extracted:

Patient characteristics: year of birth, sex, postal code of the place of residence, death date (day/month), if applicable,

chronic diseases (according to the French "Affections Longue Durée" [ALD] classification of chronic diseases), and type of health insurance coverage.

- *Outpatient information*: consultations at hospital outpatient clinics, appointments with and treatments provided by general practitioners (GPs), nurses, physiotherapists, radiologists, and specialist physicians (with the date specified as the month and year only), and all prescription medications dispensed in community pharmacies (day/month/year). Drugs were coded according to the French CIP13 classification.
- *Inpatient care*: hospital stays were classified as "acute hospital admissions," "postacute care and rehabilitation," or "hospital at home." Information about the stays contained the diagnoses (coded according to the French version of the International Statistical Classification of Diseases and Related Health Problems, 10th Revision [ICD10]), the medical procedures (coded according to the French "Classification Commune des Actes Médicaux" [CCAM] classification), and the admission and discharge dates (day/month/year).
- Stays in psychiatric units (day/month/year of admission and discharge).
- *Other information*: high-cost medications in "acute hospital admissions" and "postacute care and rehabilitation," and medical devices in "acute hospital admissions."

The extraction was performed by the Regional Health Agency ("Agence Régionale de Santé," Hauts-de-France, France) and the source data contained 26 flat files of 1 Gb.

Description of the OMOP CDM

Detailed specifications for the OMOP CDM (version 5.0) are available online.[21] The CDM is composed of 39 tables, in six groups:

- *Standardized clinical data*: the 13 tables contain the patient's demographic data, as well as information on clinical events, that occurred during the observation period.
- *Standardized vocabularies*: all basic units of information from terminologies and vocabularies used for clinical data (e.g., ICD10, sex, and types of visits) are gathered into a single table (concept). The 11 other tables specify mapping between terminologies and relationships between items.
- *Standardized health system data*: the three tables contain information about the health care provider responsible for patient care (the institution and/or the physician).
- *Standardized health economics*: the five tables contain cost information.
- *Standardized derived elements*: the five tables' aggregate information derived from raw data and that are useful for analyses (e.g., whether or not the patient is part of a cohort, the period of exposure to a drug, or an event).
- Standardized metadata: the table contains metadata about data sources (holder, description, extract–transform–load [ETL] processes, etc.).

The Integration Process

We followed the OMOP CDM specifications on loading the data.[21–23] We kept in mind the fact that the ETL process has to be capable of integrating new data extractions. The integration process is summarized in ►Fig. 1 and detailed below.

Audit

We used WhiteRabbit (a software tool developed by the OHDSI consortium) to audit the data.[24] The software scanned source data to provide detailed information on the 26 flat files (e.g., volume) and their fields. It summarized frequencies, modalities, and missing data for each field in each file. For each field, three investigators (A.L., N.D., and J.B.B.) subsequently determined (1) the type of value (structured quantitative values, structured ordinal values, or unstructured texts), (2) the type of information to which the raw data were linked, and (3) whether the field contained missing values or outliers, and therefore had to be transformed before use in the CDM. When a field corresponded to a code, the investigators searched for the terminology used (e.g., K633 corresponds to a code from the French ICD10).

Extract–Transform–Load Specifications

We use the Rabbit-in-a-Hat graphic tool to draw up specifications for the ETL process.[24] These specifications described how data were transformed from the source data model into the OMOP CDM. The two-step ETL process comprised conceptual mapping and structural mapping. Conceptual mapping links concepts related to the French context to the OMOP's international concepts, whereas structural mapping links source fields and tables to the OMOP CDM.

Two investigators (A.L. and N.D.) characterized the French context, and suggested corresponding concepts in the OMOP vocabularies. They mapped the source files to OMOP tables, and linked source fields to OMOP fields. The source fields and OMOP fields differed with regard to several formats or standardized values (e.g., male or female sex was coded as 1 or 2 in the source field and as 8,507 or 8,532 in the OMOP field). Moreover, some OMOP fields did not exactly correspond to source fields, and so the latter had to be transformed. Two investigators (A.L. and N.D.) therefore defined all the logic rules required to compute the formats, values, and transformations from the source fields to the OMOP fields. Three other investigators (J.B.B., G.F., and E.C.) checked and considered the complex transformations.

Extract–Transform–LoadImplementation

We used Talend Open Studio for Data Integration (version 6.4) to implementation the ETL process.[25] The data were stored in a PostgreSQL 9.5.13 database[26] on a secure computer without network access and running the x86_64-pc-linux-gnu Ubuntu 5.4.0–6 operating system.[27,28] Data were mounted on an encrypted drive.

The ETL process was implemented as follows:

- The standardized vocabularies used in the OMOP CDM and formatted by the OHDSI consortium to fit the OMOP

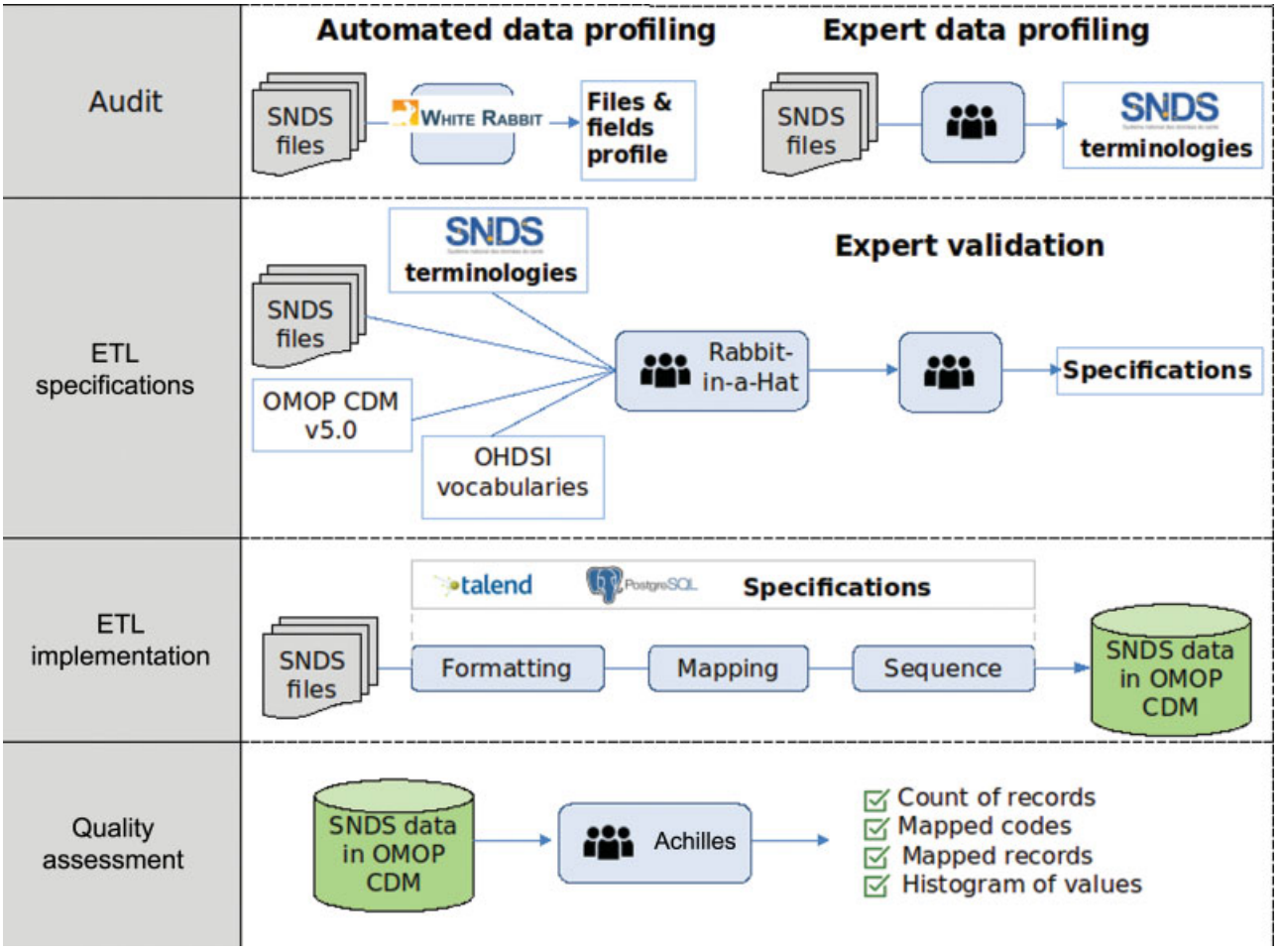

Fig. 1 A summary of the integration process. CDM, common data model; ETL, extract–transform–load; OHDSI, observational health data sciences and informatics; OMOP, observational medical outcome partnership; SNDS, Système National des Données de Santé.

CDM were downloaded through the Athena web application.[26] The SNDS terminologies, identified during the audit step (e.g., ICD10 codes), were gathered together. Several fields did not have a specific SNDS terminology, and so an investigator (A.L.) defined a specific SNDS vocabulary for these fields (e.g., sex was coded as 0, 1, or 2, corresponding to "undefined," "male," or "female"). The SNDS files were loaded with Talend into an initial database schema corresponding to the source data for the next two processes.

- SNDS fields were selected when they provided useful information from a pharmacoepidemiologic perspective. Given that SNDS fields are related to administrative data (e.g., billing information), a large number of extracted fields were not used. Relations between SNDS fields and OMOP fields (the specification step) were used to rename the SNDS fields and to change the field's format if necessary. For each field, outliers (e.g., abnormal values) were identified and excluded, according to an expert review of the audit results. Rows were selected because some patients were outside the scope of the PAERPA study (e.g., those having died before 2014). All required transformations identified in the specification step were coded with Talend grammar.
- Standardized OMOP-format clinical data were linked to standardized OMOP vocabularies, according to the OMOP specifications.

Implementation of the ETL process created a database schema corresponding to the OMOP Model. This database schema was used for all subsequent analyses.

Evaluation

In line with the OHDSI consortium's guidelines and by using the Achilles tool (H), we assess the whole ETL process.[22] For each source, file and destination table, we checked the following items: the number of records; the number of records per person (observation_period, visit_occurrence, condition_occurrence, drug_exposure, procedure_occurrence, device_exposure, and measurement). For automatic mapping, we checked the number (percentage) of mapped codes and mapped records, and the histogram of values for selected fields (demographic variables).

Results

Our assessment of the audit results and our knowledge of French healthcare databases enabled us to identify 18

Table 1 Codes and records from the SNDS mapped to the OHDSI consortium standardized vocabularies

Mapping	Number of codes (%)	Number of records (%)
UCD→ RxNorm	2,062 (100)	21,407 (100)
CIP13 → RxNorm	32,592 (95.62)	5,373,825 (88.39)
French ICD10 → ICD10 exact mapping	5,918 (85.15)	1,322,189 (83.60)
French ICD10 → ICD10 with level-1 mapping	703 (10.12)	234,427 (14.82
French ICD10 → ICD10 with level-2 mapping	325 (4.68)	24,902 (1.57%)
French ICD10 → ICD10 without mapping	4 (0.06)	21 (<0.01)

Abbreviations: CIP, Code Identifiant de Présentation, ICD10, International Statistical Classification of Diseases and Related Health Problems, 10thth Revision; OHDSI, observational health data sciences and informatics; SNDS, Système National des Données de Santé; UCD, Unité Commune de Dispensation.

vocabularies in the extracted files. We drew up specifications (with expert review) and produced an ETL procedure that complied with the OHDSI template. The document is available online.*

Several vocabularies used in French databases were added to the OMOP vocabularies: modes of admission to and discharge from the hospital units, ALD, "Activités de la Vie Quotidienne" (AVQ), CCAM, supplementary insurance status, code identifiant de présentation (CIP)13, "Catalogue spécifique des actes de rééducation et réadaption" (CSARR), sex, "Groupes Homogènes de Malades" (GHM), "Groupes Médico-Economiques" (GME), the indication for "hospital at home" care, French ICD10, "liste des produits et prestations" (LPP), the provider's specialty, consultation type, type of health care resource consumption, and "Unité Commune de Dispensation" (UCD). The "Groupes Homogènes de Séjours" (GHS) vocabulary was not added, as it is dedicated to payment activities alone. All these abbreviations are defined in ►**Supplementary Appendix A** (available in the online version).

Conceptual Mapping

Three French terminologies (CIP13, the French ICD10, and UCD) and two additional vocabularies (sex and the type of health care resource consumption) were mapped to OMOP standardized vocabularies.

The French ICD10 vocabulary was automatically mapped to ICD10 using concept code. As this process included SNDS-specific changes (notably when the French code was more precise than the ICD10 code), the code was truncated to correspond to the most precise international code. For example, W11.38 ("Chute sur ou d'une échelle," "lieu de sport," "en participant à d'autres activités précisées") in the French ICD10 was linked to W11 (fall on and from a ladder) in the international ICD10, as it was the most precise code. A total of 1,028 codes (14.79%) were associated with loss of information because they were mapped to a less precise ICD10 code; these codes corresponded to 259,329 records (16.40% of the total). Only four codes (0.06%) corresponding to 21 records (<0.01%) were not mapped.

The CIP13 and UCD vocabularies were automatically mapped to the anatomical therapeutic chemical (ATC) classification at the clinical level (using French correspondence tables) and then to RxNorm (using relationship tables provided by the OHDSI consortium). Although homeopathic and parapharmacy products were included in CIP13 and maintained in the extraction, they could not be linked to ATC and RxNorm as these terminologies only include drug compounds. As a result, 32,592 CIP13 codes (95.62% of the total number of codes in CIP13, corresponding to 5,373,825 drug administration records [88.39% of the total number of records]) could be mapped to standardized vocabularies. For all mappings to the ATC classification, French information about the drug formulation and dose level were lost because the ATC contains information about the active compound only.

►**Table 1** summarizes the outcome of the automatic mapping process for CIP13, UCD and French ICD10 codes. In line with the OMOP's specifications, we also loaded new concepts into the relationship and concept_relationship tables for the mapped codes.

In the OMOP CDM, *X_type_concept_id* variables correspond to the type of source. In the present study, these variables were mapped to OHDSI concepts by taking account of specific context of care in France, and the data collection process (►**Supplementary Appendix B**, available in the online version). Given that all the patients were covered by French compulsory health insurance, the observation period corresponded to the "period while enrolled in insurance" concept (*observation_type_concept_id*). Since data on a patient's death come from a specific register (the CepiDC), this information was mapped to the "death certificate immediate cause" concept (*death_type_concept_id*). Visits were classified as "visit derived from encounter on claim" (*visit_type_concept_id*). Procedures were characterized as "primary procedure" or "secondary procedure," depending on the clinical impact of the procedure (*procedure_type_concept_id*). Drug exposures were mapped to the "drug inpatient administration" concept when administered in hospital and to "prescription dispensed in pharmacy" when dispensed in community pharmacies (*drug_type_concept_id*). All device exposures were mapped to the "inferred from procedure claim" concept (*device_type_concept_id*). Conditions were characterized as "primary conditions" or "secondary conditions," depending on the clinical impact of the condition (*condition_type_concept_id*). Measurements were all characterized by

* https://subversion.univ-lille2.fr/gitlab/paerpa/etl_paerpa/.

the "from physical examination" concept, since they were all performed by physicians (*measurement_type_concept_id*).

Twelve source vocabularies were not mapped to OHDSI vocabularies. However, the records were still been loaded into the OMOP CDM and could be manipulated though the *X_source_concept_id* column, which references the source.

Structural Mapping

We assessed all the files/entities extracted from the SNDS. We loaded 11 tables from the OMOP CDM's "standardized clinical data" section, and the three tables from the "standardized health system data" section. We did not load the "observation," "specimen" and "note" tables, since our SNDS extraction does not provide information about laboratory results or consultation reports. The "standardized health economics" tables were not loaded. Condition_era and drug_era were loaded with the script provided by the OHDSI consortium. These transformations are illustrated in ►Fig. 2.

The main structural transformations consisted in (1) merging different sources for the same entity (visits, procedures, diagnoses, and drugs), (2) replicating data when a row contains several items of information, and (3) concatenating fields when the source primary key is a combination of several source fields (the OMOP data model restricts the use of a single field for a unique identifier).

Some items of information were incomplete in the French health care system, and so prompted the following alterations: (1) the exclusion of records from 109 sets of twins (218 patients) because the SNDS does not allow distinct records for twins, and (2) imputation of default values when the day was not provided in the SNDS; if only the month and year of an outpatient stay were specified, the day was imputed as the first day of the month.

Integration Results

The numbers of records loaded into each table in the OMOP CDM are summarized in ►Table 2. The median (interquartile range [IQR]) per patient was 160 (19; 364) for inpatient and outpatient stays, 30 (3; 55) for consultations with a family physician, 2 (0; 6) for inpatient stays. The Achilles tool was used to detect data quality problems, such as an undefined identifier (e.g., person_id, care_site_id), incorrect values (e.g., an observation period end date that occurred before the start date), and the occurrence of procedures, diagnoses or drug administrations outside the observation period. Moreover, Achilles' provided valuable summary information about our population and the available data per person (►Figs. 3 and 4, available in the online version). In addition to performing a descriptive statistical analysis, we were able to assess several usage examples related to the PAERPA project (detailed in ►Supplementary Appendix C [available in the online version] and ►Fig. 5).

Discussion

The present study comprised the structural and conceptual mapping of French medical and administrative data against the OMOP CDM. To this end, we developed audit procedures, specifications, ETL procedures, and methods for evaluating the final database. This work was performed as part of the PAERPA project that provided comprehensive data on outpatient and inpatient visits over 4 years, for around 38,730 patients.

The main strength of the present study was its extraction of data from different sources and with different durations (e.g., acute inpatient stays, rehabilitation inpatient stays, and outpatient visits) via a unique list of concepts. Integration of the

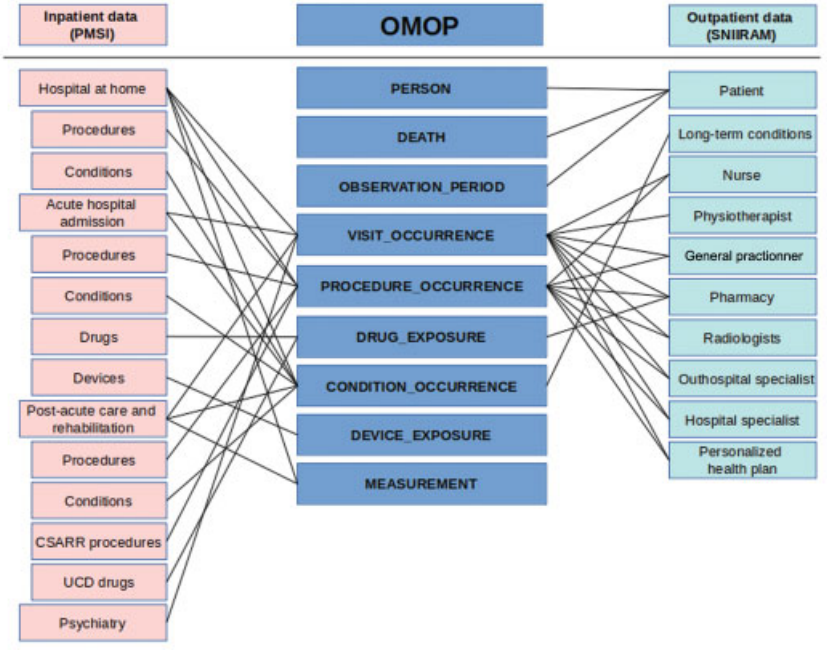

Fig. 2 A schematic diagram of the structural mapping between our extraction of the SNDS data and the OMOP CDM. Hospital-sourced data are shown in red, outpatient data are shown in green, and the final OMOP tables are shown in blue. PMSI, Program de Médicalisation du Système d'Information; SNIIRAM.

Table 2 Number of records integrated from the SNDS into OMOP tables, and corresponding numbers per patient

SNDS Tables	Number of records	OMOP Table	Number of records	Median (IQR) number per patient
Patient	38,730	Person	38,730	–
–	–	Observation_period	38,730	1 (1–1)
Inpatient and outpatient stays	15,678,382	Visit_occurrence	15,678,382	160 (19–364)
Family physician	1,321,486	Family physician	1,321,486	30 (3–55)
Nurse	9,394,009	Nurse	9,394,009	14 (0–85)
Inpatient	237,675	Inpatient	237,675	2 (0–6)
Inpatient and outpatient conditions	1,667,8457	condition_occurrence	16,678,457	11 (2–40)
Inpatient and outpatient drugs	6,100,837	drug_exposure	6,100,837	125 (0–257)
Inpatient and outpatient procedures	32,817,285	procedure_occurrence	32,817,285	368 (82–822)
Devices	29,554	device_exposure	29,554	0 (0–0)
		measurement	464,301	0 (0–0)
–		drug_era	811,334	14 (0–33)
–		condition_era	151,106	3 (2–5)

Abbreviations: OMOP, observational medical outcome partnership; SNDS, Système National des Données de Santé.

data into the OMOP CDM revealed two main advantages: (1) data standardization avoids the need to handle data in different formats (particularly for data recorded in outpatient and inpatient contexts), and (2) the data were easier to query (avoiding multiple join operations) because similar concepts initially stored in different tables were gathered together.

Other major studies performed in Germany, the United States, the United Kingdom, Brazil, China, and Korea have focused on the integration of hospital databases[14,15,17,29–32] or a particular segment of outpatient care (e.g., primary care[33,34]). There appear to be more concepts in the French sources than in the implementations of OMOP CDM in these

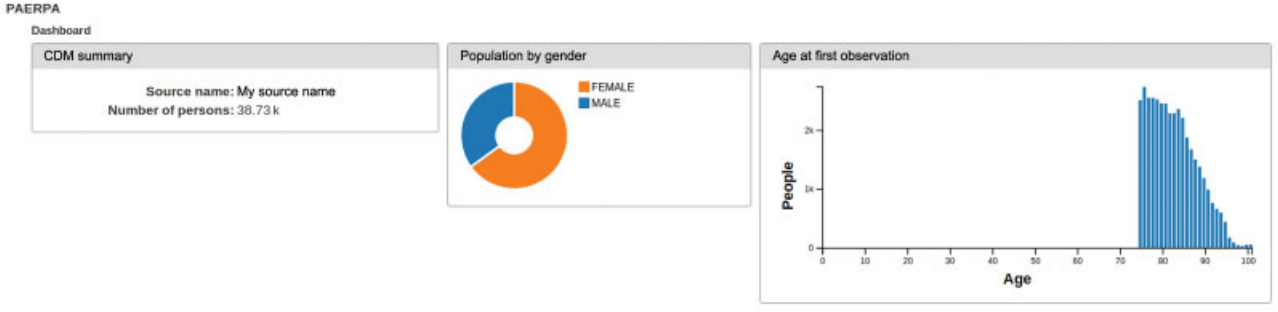

Fig. 3 Dashboard summary generated by Achilles. CDM, common data model; PAERPA, Personnes Agées en Risque de Perte d'Autonomie.

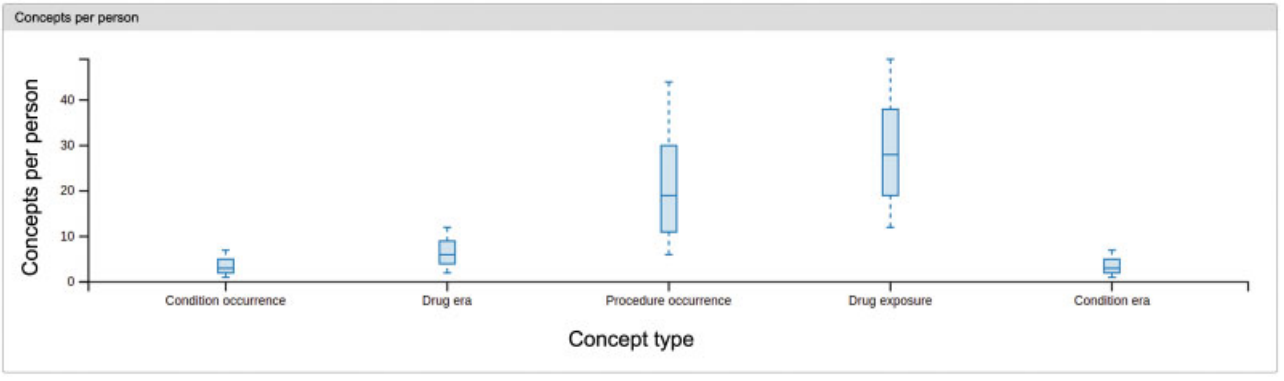

Fig. 4 Concept type per person computed by Achilles.

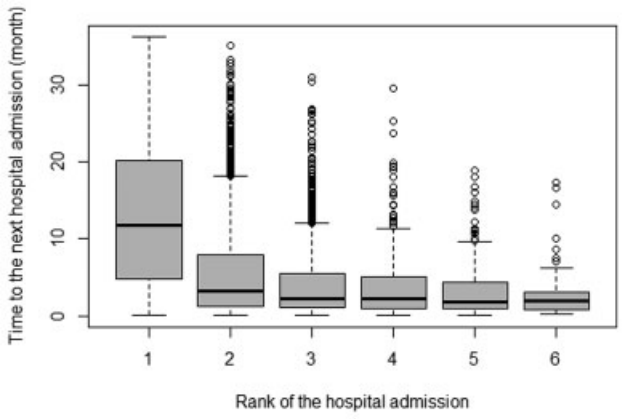

Fig. 5 Time to the next hospital admission according to the rank of the hospital admission (also see ►**Appendix C**, available in the online version).

countries, since the French health care system covers both inpatient and outpatient care for the majority of the French population. For the conceptual mapping of visits (the *visit_occurrence* table), for example, the corresponding data were sourced from four inpatient files and eight outpatient files.

The European Health Data and Evidence Network (EHDEN) is seeking to integrate more than 100 million health records into the OMOP CDM (G). The centralized French health care system is a valuable resource in this respect because it provides nationwide health information on more than 65 million people.

One of the main challenges in the present work was adapting data recorded in the French cultural context to match the OMOP's international concepts. As reported by Garza et al, the loss of information about the context of use is a concern when standardizing data.[8] The use of local vocabularies was also an obstacle to conceptual mapping, since as French health care system does not use international vocabularies other than an adaptation of ICD10.

Some French vocabularies (e.g., CCAM, CSARR, and LPP) will still have to be mapped to OMOP vocabularies, to (1) fully benefit from the advantages of a CDM, and (2) compare the results from different countries. The mapping of local vocabularies to international vocabularies requires medical, IT, and linguistic skills, and (to ensure consistency) regular interaction with the teams in charge of target terminology. Fortunately, the fact that diagnoses and drugs have been mapped to reference terminologies in OMOP will allow us to use the statistical methods developed by the OHDSI consortium to conduct pharmacoepidemiologic studies.

Other major difficulties were related to the source data (e.g., erroneous drug codes, a missing provider identifier, or imprecise visit dates due to anonymization of the source data). To overcome these problems, OMOP CDM provides fields for characterizing source data. These fields are used to store concept_id of the local vocabularies and concepts corresponding to the local context (e.g., the type of death certificate), which avoids the loss of source information. In contrast, the OHDSI consortium encourages the development of new concepts related to the local context when existing vocabularies do not fully document the facts. These new concepts should be developed within the OHDSI consortium, to maintain consistency with other developments and to benefit from the community's experience. Hence, we have suggested concepts related to the French context (e.g., types of interaction with healthcare providers in France).

Another limitation of our study is that only a sample of the French health care system was integrated. Nevertheless, several aspects lead us to expect that integration of the whole database will be possible. First, our study was based on conventional SNDS fields (hospital stays, medications, etc.) that are likely to be extrapolated. Second, data volumetry in the SNDS could be addressed by selecting only the relevant information in the three databases, and by excluding fields relating to accounting information. Third, for pharmacoepidemiologic research, our data model will only need to be fed annually (rather than daily) but could be queried several times a year.

A final limitation concerns the use of the OMOP CDM in version 5.0. Since the beginning of our project, version 6 has been released. The main developments concern (1) the details of the stay, (2) the information extracted from the text notes with natural language processing techniques, and (3) the information from the questionnaire. These changes have no impact on our results since the SNDS data do not include free text and questionnaires, and VISIT_OCCURRENCE was populated by all the information about the hospital stay (for hospital stay). In addition, the R packages, developed by the OHDSI community, are still available for OMOP v5.0.

Our work enables the analyses of these data using the OHDSI consortium statistical framework. We consider that analysis of a patient population of this size constitutes a unique opportunity to conduct pharmacoepidemiologic studies and active postmarket drug safety surveillance in France. One of the first steps will be the calibration and reproduction of previous pharmacoepidemiologic studies in the French context. Furthermore, this semantic interoperability would enable international databases to be built, giving the statistical power required to detect very rare adverse drug reactions.

Conclusion

Our present results demonstrate that data from the French national health care system can be integrated into the OMOP CDM. This work provides access to the advantages of a CDM, namely, data standardization and use of tools developed by the research community.

Clinical Relevance Statement

The Observational Medical Outcome Partnership's common data model (OMOP CDM) provides analytical tools for active postmarket drug safety surveillance via electronic health records (EHRs). We assessed the feasibility of transforming the nationwide French EHR database into the OMOP CDM's format.

Multiple Choice Questions

1. Into which common data model were our data integrated:
   a. Sentinel
   b. PCORnet
   c. OMOP
   d. CDISC-SDTM

   Correct Answer: The correct answer is option c. In this study, we choose to integrate data from the French health care system into the OMOP common data model.

2. What is the main challenge when integrating data into a common data model?
   a. Dealing with data volume
   b. Taking account of the context in which the source data were recorded
   c. Master database technologies
   d. Vendor demos

   Correct Answer: The correct answer is option b. The main challenge is to keep information about the context in which data were recorded, for example the specificities of the health care system.

Protection of Human and Animal Subjects

Human or animal subjects were not included in the project.

Funding

This work is part of the "PAERPA: Health Pathway of Seniors for Preserved Autonomy" project funded by the Agence Régionale de Santé (ARS) Hauts-de-France. The ARS Hauts-de-France provided the data and funding. The funding bodies did not have an influence on study design and data analysis but were involved in drafting the manuscript.

Conflict of Interest

G.F. reports grants from ARS Hauts-de-France, during the conduct of the study. E.C. reports grants from ARS Hauts-de-France, during the conduct of the study. A.L. reports grants from ARS Hauts-de-France, during the conduct of the study. M.M.D. reports grants from ARS Hauts-de-France, during the conduct of the study. J.B.B. reports grants from ARS Hauts-de-France, during the conduct of the study. M.D. reports personal fees from French Ministry of Health, (DREES), during the conduct of the study. N.D, reports grants from ARS Hauts-de-France, during the conduct of the study.

Acknowledgments

We thank Mariama Diallo and Dr. Alexandre Caron for their invaluable assistance with the ETL process. We also thank Nicolas Paris and Anaïs Payen for their invaluable helps and constructive advices.

References

1 Sentinel Initiativetp. Available at: www.mini-sentinel.org/. Accessed April 4, 2019
2 National Patient-Centered Clinical Research Network. PCORnet. Available at: https://pcornet.org/pcornet-common-data-model/. Accessed April 4, 2019
3 Chazard E, Merlin B, Ficheur G, Sarfati J-C, Beuscart R; PSIP Consortium. Detection of adverse drug events: proposal of a data model. Stud Health Technol Inform 2009;148:63–74
4 i2b2: Informatics for Integrating Biology & the Bedside. Available at: https://www.i2b2.org/. Accessed April 4, 2019
5 OHDSI—Observational Health Data Sciences and Informatics. Available at: https://www.ohdsi.org/. Accessed April 4, 2019
6 Weeks J, Pardee R. Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in U.S. health care research. EGEMS Wash DC 2019;7(01):4
7 Liyanage H, Liaw S-T, Jonnagaddala J, Hinton W, de Lusignan S. Common data models (CDMs) to enhance international big data analytics: a diabetes use case to compare three CDMs. Stud Health Technol Inform 2018;255:60–64
8 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
9 CDISC. Study Data Tabulation Model (SDTM). Available at: https://www.cdisc.org/standards/foundational/sdtm. Accessed April 4, 2019
10 Xu Y, Zhou X, Suehs BT, et al. A comparative assessment of observational medical outcomes partnership and mini-sentinel common data models and analytics: implications for active drug safety surveillance. Drug Saf 2015;38(08):749–765
11 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
12 Ryan PB, Stang PE, Overhage JM, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf 2013;36(Suppl 1):S143–S158
13 Schuemie MJ, Gini R, Coloma PM, et al. Replication of the OMOP experiment in Europe: evaluating methods for risk identification in electronic health record databases. Drug Saf 2013;36(Suppl 1): S159–S169
14 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
15 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
16 You SC, Lee S, Cho S-Y, et al. Conversion of national health insurance service-national sample cohort (NHIS-NSC) database into observational medical outcomes partnership-common data model (OMOP-CDM). Stud Health Technol Inform 2017;245:467–470
17 Yoon D, Ahn EK, Park MY, 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(01):54–58
18 Le dispositif Paerpa. Ministère des Solidarités et de la Santé. Available at: https://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/parcours-des-patients-et-des-usagers/le-parcours-sante-des-aines-paerpa/article/le-dispositif-paerpa. Accessed April 4, 2019
19 Tuppin P, de Roquefeuil L, Weill A, Ricordeau P, Merlière Y. French national health insurance information system and the permanent beneficiaries sample. Rev Epidemiol Sante Publique 2010;58(04): 286–290
20 Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc J-L, Sailler L. French health insurance databases: what interest for medical research? Rev Med Interne 2015;36(06):411–417
21 OHDSI/Common Data Model. Definition and DDLs for the OMOP Common Data Model (CDM). Available at: https://github.com/OHDSI/CommonDataModel. Accessed April 4, 2019
22 Observational Health Data Sciences and Informatics. ELT creation best practices. Available at: http://www.ohdsi.org/web/wiki/doku.php?id=documentation:etl_best_practices. Accessed April 4, 2019

23 OHDSI/Achilles. Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) - descriptive statistics about a OMOP CDM database; 2019. Available at: https://github.com/OHDSI/Achilles. Accessed October 17, 2019
24 Observational Health Data Sciences and Informatics. WhiteRabbit. Available at: http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:whiterabbit. Accessed April 4, 2019
25 Open Source ETL—Talend Open Studio for Data Integration. Talend real-time open source data integration software. Available at: https://www.talend.com/products/data-integration/data-integration-open-studio/. Accessed April 4, 2019
26 Postgre SQL. The world's most advanced open source database. Available at: https://www.postgresql.org/. Accessed April 4, 2019
27 Ubuntu. The leading operating system for PCs, IoT devices, servers and the cloud. Available at: https://www.ubuntu.com/. Accessed April 4, 2019
28 Athena. Available at: http://athena.ohdsi.org/search-terms/terms. Accessed April 4, 2019
29 Lima DM, Rodrigues-Jr JF, Traina AJM, Pires FA, Gutierrez MA. Transforming two decades of ePR data to OMOP CDM for clinical research. Stud Health Technol Inform 2019;264:233–237
30 FitzHenry F, Resnic FS, Robbins SL, et al. Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. Appl Clin Inform 2015;6(03): 536–547
31 Zhang X, Wang L, Miao S, et al. Analysis of treatment pathways for three chronic diseases using OMOP CDM. J Med Syst 2018;42(12): 260
32 Viernes B, Lynch KE, South B, Coronado G, DuVall SL. Characterizing VA users with the OMOP common data model. Stud Health Technol Inform 2019;264:1614–1615
33 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
34 European Health Data Evidence Network Available at: https://www.ehden.eu/. Accessed October 9, 2019

Erratum: Article has been corrected as per Erratum published on March 11, 2020. DOI of the Erratum is 10.1055/s-0040-1702166.