




Research and Applications

Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned

Marek Oja , PhD^{1,*}, Sirli Tamm, MSc¹, Kerli Mooses, PhD¹, Maarja Pajusalu, MSc¹, Harry-Anton Talvik, MSc^{1,2}, Anne Ott, MSc¹, Marianna Laht, MD¹, Maria Malk, MSc¹, Marcus Lõo, MSc¹, Johannes Holm, MSc¹, Markus Haug, MSc¹, Hendrik Šuvalov, MSc¹, Dage Särg, MSc^{1,2}, Jaak Vilo , PhD^{1,2}, Sven Laur, PhD¹, Raivo Kolde, PhD¹, Sulev Reisberg , PhD^{1,2}

¹Institute of Computer Science, University of Tartu, 51009 Tartu, Estonia, ²STACC, 51009 Tartu, Estonia

*Corresponding author: Marek Oja, PhD, Institute of Computer Science, University of Tartu, Narva mnt 18, 51009 Tartu, Estonia (marek.oja@ut.ee)

Author Contributions: Dr M. Oja and S. Tamm are considered co-first authors and Dr R Kolde and Dr S Reisberg are considered co-last authors of this work. In addition, they had full access to all the data in the study and take responsibility for the integrity of data and accuracy of the data analysis.

Abstract

Objective: To describe the reusable transformation process of electronic health records (EHR), claims, and prescriptions data into Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM), together with challenges faced and solutions implemented.

Materials and Methods: We used Estonian national health databases that store almost all residents' claims, prescriptions, and EHR records. To develop and demonstrate the transformation process of Estonian health data to OMOP CDM, we used a 10% random sample of the Estonian population ($n=150\,824$ patients) from 2012 to 2019 (MAITT dataset). For the sample, complete information from all 3 databases was converted to OMOP CDM version 5.3. The validation was performed using open-source tools.

Results: In total, we transformed over 100 million entries to standard concepts using standard OMOP vocabularies with the average mapping rate 95%. For conditions, observations, drugs, and measurements, the mapping rate was over 90%. In most cases, SNOMED Clinical Terms were used as the target vocabulary.

Discussion: During the transformation process, we encountered several challenges, which are described in detail with concrete examples and solutions.

Conclusion: For a representative 10% random sample, we successfully transferred complete records from 3 national health databases to OMOP CDM and created a reusable transformation process. Our work helps future researchers to transform linked databases into OMOP CDM more efficiently, ultimately leading to better real-world evidence.

Lay Summary

Health data can be found in various sources and formats, making it challenging for researchers. To address this issue, one possible approach is to transform the data into a standardized common data model (CDM). In this study, we describe the process of converting electronic health records (EHR), claims, and prescriptions data into the Observational Medical Outcome Partnership (OMOP) CDM, along with the challenges faced and solutions implemented. We used Estonian national health databases containing information on claims, prescriptions, and EHR records of 10% of Estonian residents (MAITT dataset). The study describes how data were mapped to standardized vocabulary and successfully converted to the OMOP CDM. We discuss the encountered difficulties and problems and propose solutions to help future researchers transform linked databases into OMOP CDM more efficiently, leading to better real-world evidence.

Key words: OMOP; electronic health record; EHR; ETL; mapping.

Background and significance

While randomized controlled trials are the gold standard for causative clinical studies, generating real-world evidence (RWE) from routinely collected real-world health data (RWD) has gained more and more attention in recent years as it provides information about a broader patient population in a less controlled environment when compared to the clinical trials, and it better reflects what is actually happening in the clinical practice.^{1–3} RWD can be used for a large variety of studies—for example, for the scientific evaluation of usage,

effects, potential benefits, and risks of a medical product,³ drug adherence, health and treatment patterns, treatment guidelines.⁴

To generate high-quality RWE, diverse RWD from different healthcare settings and geographical locations integrated into data networks are needed.¹ Still, linking different real-world datasets and conducting multidatabase studies to produce high-quality RWE is a challenging task. It has been shown that many of the problems could be solved by using a single CDM for all datasets.^{5–7} This would improve not only

Received: March 24, 2023; Editorial Decision: October 24, 2023; Accepted: November 15, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of the American Medical Informatics Association.

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

the quality of the outcome but also their acceptability in decision-making.⁶ In recent years, there has been an increasing interest in transferring health data to the OMOP CDM,^{8,9} which offers a standardized vocabulary and structure, improving the interoperability between databases. Moreover, several open-source software solutions have been developed to support the transformation and analysis process.¹⁰ This all supports the transformation process of health data to OMOP and could be considered one key reason why OMOP has become increasingly popular.

Previous research has described the successful transformation of data to OMOP CDM, which originate from different sources like biobanks,¹¹ national databases, and registries,^{12–14} hospital databases,^{15–18} questionnaires,¹⁹ cohort studies.^{20–22} Some studies focus on specific conditions or some part of a database^{12,13,16,17,19,21} while others transfer whole databases with different diagnoses, drug adherence, or health care procedures.^{11,15,20} Despite the existing research, it has been stressed that continued sharing of experiences, methodologies, and challenges of the data transformation process to OMOP is needed as it helps to develop the transformation process and foster collaboration.^{21,23}

Today OHDSI network includes more than 453 databases mapped to the OMOP CDM,¹⁰ and an initiative to establish a federated network of OMOP healthcare datasets across Europe has been coordinated and partly funded in the European Health Data & Evidence Network (EHDEN) project.²⁴ However, the geographic distribution of OMOP datasets is uneven as the real-world datasets from Eastern European countries are much less represented than Western countries, leading to gaps in RWE from these regions.²⁵ In addition, the number of datasets that contain data from several healthcare settings is small. However, in most clinical and epidemiological studies, information from electronic health records (EHR), claims, pharmacies, etc., is needed.

To the best of our knowledge, there is a lack of studies describing the integration process of claims data, EHRs, and prescriptions data into a single complete patient-centered view. In Estonia, 3 separate national electronic health databases store such information separately. These databases use different coding systems and structure which have hindered the co-use of these databases. To address these issues, we developed a reusable process to transform these 3 separate electronic health databases into a single coherent patient-centric OMOP CDM. The current article describes the transformation process, challenges faced, and solutions implemented.

Methods

Data sources

Estonia is a small country in Northern-Eastern Europe with a population of 1.3 million consisting primarily of Estonians (70%). It is mandatory for all healthcare providers in Estonia to use 3 central operational health databases to enable easy data exchange between the institutions and interoperability. These databases cover clinical information from almost all healthcare settings (hospitals, specialists, family doctors, labs, pharmacies). Data could be potentially linked using personal identification codes provided to all residents. The main content and terminologies used in the databases are described in Table 1. There is some difference between the databases in the data coverage. EHRs store data from all private and state-owned healthcare providers for insured and uninsured

people, while health insurance claims include about 95% of the Estonian population who have public insurance.²⁶ Using the claims database is mandatory for reimbursement. All prescription drugs are prescribed digitally and stored in the corresponding database.²⁷ None of the health datasets is the primary source for death information (we were not allowed to use death registry data in this work), containing deaths related to healthcare services only (67%).

Data from these 3 health databases complement each other; thus, linking adds additional value and improves the data quality. For example, health insurance claims provide information on which services (eg, lab tests) have been provided, while further details of these services (eg, the results of the tests) can be found in EHR. The drug prescription database does not contain information for over-the-counter medications where no prescription is needed (eg, paracetamol) or inpatient medications, but such information can be sometimes found in free text in EHR. At the same time, diagnoses, drugs, and procedures may be recorded in multiple source datasets, which adds complexity to the linking of health data. The drawback of EHR is its lower quality as the information is partially recorded in a semistructured or free-text format (eg, the weight or blood pressure of the patient is given in clinical notes) which is challenging to use in automated analysis. Also, case summaries in EHR contain only the most relevant information but may miss other tests or services conducted on a patient.

Carrying out an epidemiological study in Estonia requires obtaining approval from an ethics committee, followed by the collection of necessary data, typically from all 3 databases. To develop and demonstrate the transformation process of Estonian health data to OMOP CDM, we used a 10% random sample of the Estonian population ($n=150\,824$ patients) from 2012 to 2019 (MAITT dataset) (Figure 1). For the sample, complete information from all 3 databases was extracted. This work was approved by the Estonian Bioethics and Human Research Council (EBIN, no. 1.1-12/653).

Extract, transform, and load process

We transformed source data to the OMOP CDM version 5.3, which includes 15 clinical data tables for storing patient demographics and clinically relevant information. The transformation process had 3 main stages: first, creating the mapping between source and OMOP vocabularies; second, coming up with the technical implementation around the mapping process, starting with data extraction from source databases and ending with loading the data into the target database (extract, transform, load, ETL); and finally, validating the transformation results (Figure 2).

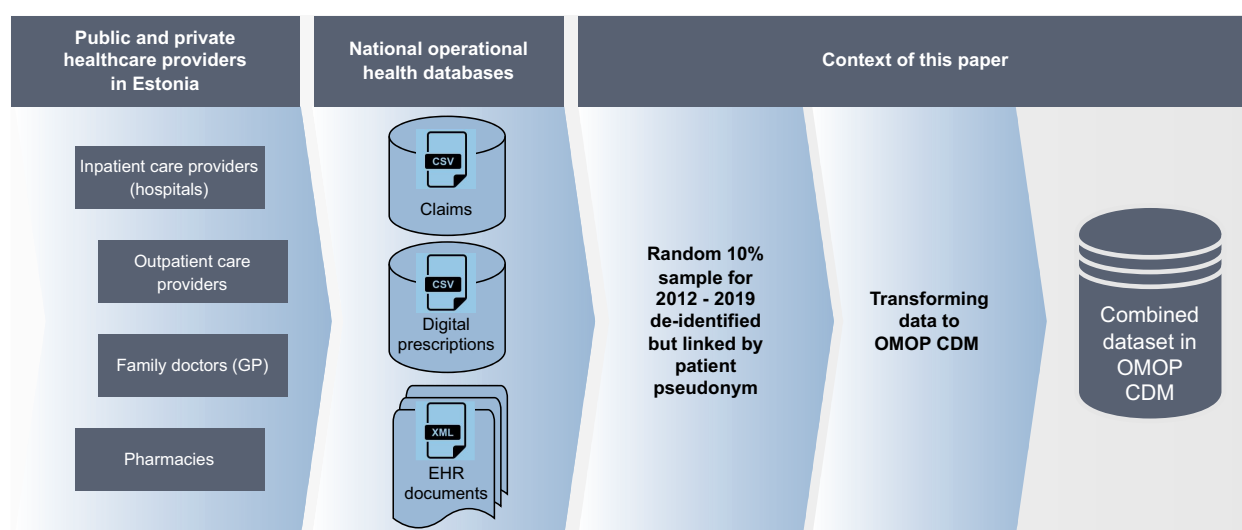
Mapping source vocabularies to OMOP vocabularies

Three source datasets (EHR, prescriptions, claims) used different terminologies to represent clinical events, and thus, they were all mapped to standard OMOP vocabularies. The standard vocabularies which are used in OMOP for medical information are SNOMED Clinical Terms (CT), RxNorm,²⁸ Logical Observation Identifiers Names and Codes (LOINC), Unified Code for Units of Measure (UCUM), and OHDSI community-developed vocabularies (Cancer Modifier, OMOP Extension, and RxNorm Extension). As for nonstandard vocabularies, 2 approaches were used. Where mappings between nonstandard and standard vocabularies (eg, from International Classification of Diseases 10th revision ICD-10

Table 1. National health databases in Estonia.

Data source	Content	Terminologies
Electronic health records (EHR)	<ul style="list-style-type: none"> • inpatient and outpatient case summaries (including medical history, lab tests, and their results, procedures, drugs, and allergies) • referrals and their responses 	<ul style="list-style-type: none"> • ICD-10 diagnosis codes • SNOMED CT or local codes for procedures and results • LOINC codes for lab tests • TNM codes and cancer stages • Other terminologies used in free text parts of clinical notes
Drug prescriptions	<ul style="list-style-type: none"> • product name and code • ATC code • amount • administration guidelines • purchase date and location • the healthcare provider who issued the drug • ICD-10 diagnosis 	<ul style="list-style-type: none"> • ATC codes • Local drug product codes • ICD-10 diagnosis codes
Health insurance claims	<ul style="list-style-type: none"> • diagnosis codes • services provided (eg, admissions, lab tests, procedures, drug administrations) • surgical procedures conducted 	<ul style="list-style-type: none"> • ICD-10 diagnosis codes • Local service codes • NCSP codes

Abbreviations: ATC, Anatomical Therapeutic Chemical classifications; ICD-10, International Classification of Diseases 10th revision; LOINC, Logical Observation Identifiers Names and Codes; NCSP, Nordic Medico-Statistical Committee Classification of Surgical Procedures; SNOMED CT, SNOMED Clinical Terms; TNM, Classification of Malignant Tumors.

**Figure 1.** The data acquisition process of national health databases in Estonia and the context of this article.

to SNOMED CT) already existed in the OHDSI vocabularies repository Athena,²⁹ the existing mappings were used. Otherwise, manual mapping was performed, prioritizing the most frequent and essential terms for the planned research studies. For manual mapping, we used the Usagi³⁰ tool, an application provided by the OHDSI community to help create mappings between nonstandard source concepts and standard concepts. Manual mappings were validated by medical specialists. Next, we give a more detailed overview of the mapping process based on different source codes in our databases.

The diagnosis codes in all 3 data sources were in ICD-10 format and were mapped to the OMOP vocabulary using the available mappings in Athena.

Lab tests from EHR were encoded using LOINC codes. As LOINC codes are standard in the OMOP vocabulary, no additional mapping was required. However, we needed additional mappings for lab test results. In cases where lab test

results were presented as text (eg, “negative” or “positive”), these were standardized, and the appropriate answer was determined based on the corresponding LOINC code. If a LOINC answer code was not present in the OMOP LOINC vocabulary or was determined to be invalid, a mapping to SNOMED CT was created.

To convert Anatomical Therapeutic Chemical classifications (ATC) codes to RxNorm, we used the standard ATC-to-RxNorm mapping available in Athena. However, as ATC codes provide information at the ingredient level and do not include information about strength or drug form, we also used Estonian-specific drug product codes to manually map source data to the RxNorm clinical drug level (ingredient, their strength, and form). If a drug product code could not be mapped, the ATC-to-RxNorm mapping was used instead.

The service codes are used on claims for administrative purposes. These are Estonian-specific codes, and the previous

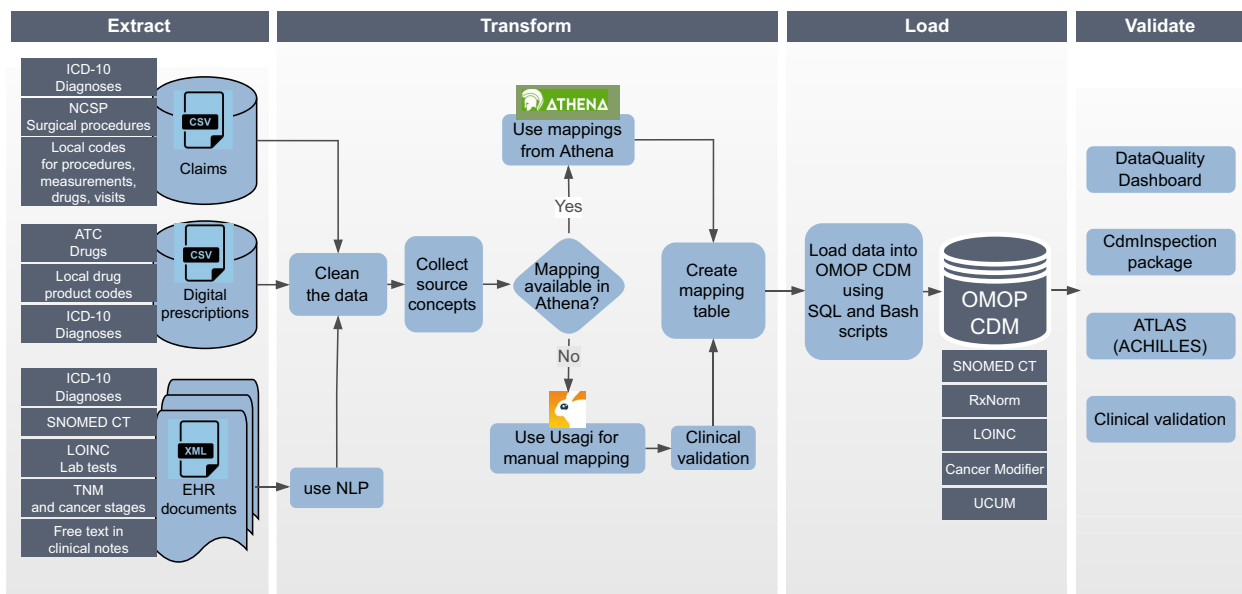


Figure 2. Overview of the data transformation process.

mapping to OMOP vocabularies was unavailable. These codes are divided into subgroups which include similar services, such as visits, procedures, surgeries, measurements without results, and drugs (eg, specific cancer or hospital-given drugs). To map these codes to OMOP vocabularies, the codes were first translated to English using publicly available Google Translate service, and then the Usagi tool was used to map these codes to valid SNOMED CT, RxNorm, or LOINC codes, depending on the nature of the source code.

Similarly to service codes, mappings for Nordic Medico-Statistical Committee Classification of Surgical Procedures (NCSP) codes and cancer-specific findings (TNM classification of malignant tumors, cancer stages, histopathology grades) were created using Usagi. The NCSP codes were mapped to standard SNOMED CT codes, and cancer-specific findings were mapped to the Cancer Modifier vocabulary.

If the source code could not be mapped to OMOP standard concept, the record was transferred to OMOP CDM with a concept identifier of “0” (meaning no matching concept) and placed in the appropriate target table based on the nature of the source concept. With this approach, the source concepts can still be used for defining studies within this dataset, however, such definitions likely will not work elsewhere.

Technical implementation

The ETL process was built using a combination of Bash scripts, SQL scripts, and manual comma-separated values mapping files. The ETL process is shown in Figure 2. The process began with cleaning and validating the source data. From EHR documents, the necessary information was extracted from structured files of Extensible Markup Language and cleaned. As the format of the information is partially free-text or semistructured in the EHR documents, natural language processing methods were used to extract this information. The data from the 3 source datasets were combined into a single OMOP CDM database using developed pipelines and mappings. As there is no direct link between the same event in different data sources, no duplicate removal was performed. For example, if the same diagnosis

code comes from an EHR document and claims, both are converted to OMOP CDM. Still, we determined the provenance of the information, whether it came from an EHR document, claim, or prescription data. The developed ETL process is reusable for other samples on these data sources.

Validation

The results of the ETL process were assessed using an open-source Achilles data characterization tool,³¹ DataQualityDashboard version 1.4.1,⁷ and the EHDEN CdmInspection tool.³² Achilles data characterization tool³¹ allows getting an overview of the converted data. DataQualityDashboard⁷ runs over 3000 checks on conformance, completeness, and plausibility on the data transformed to OMOP CDM. CdmInspection tool³² runs additional vocabulary (eg, top unmapped and mapped codes in different tables) and infrastructure checks compared to Achilles and DataQualityDashboard. Any errors, warnings, or issues found were addressed by revising the ETL code or mappings, and the process was repeated until all solvable errors were resolved.

Results

The transformation process for the OMOP CDM was developed using a sample of 10% of the Estonian population ($n = 150\,824$ patients) with a similar age and gender distribution as the overall population (Figure 3). All persons have one observation period covering January 1, 2012 to December 31, 2019 except when the person was born later than 2012 or died before 2019. In these cases, the observation period was adjusted accordingly. In total, the sample dataset contained 4 970 022 EHR documents, 6 222 818 claims, and 9 289 527 digital prescriptions.

Out of 150 824 persons in the source data, we were able to transform 149 364 (99.0%) persons and their medical data to OMOP CDM. The remaining 1460 persons did not have a birth year reported in any source datasets, which is a mandatory attribute in OMOP CDM. These patients had altogether 2364 entries in source data. On the other hand, linking 3

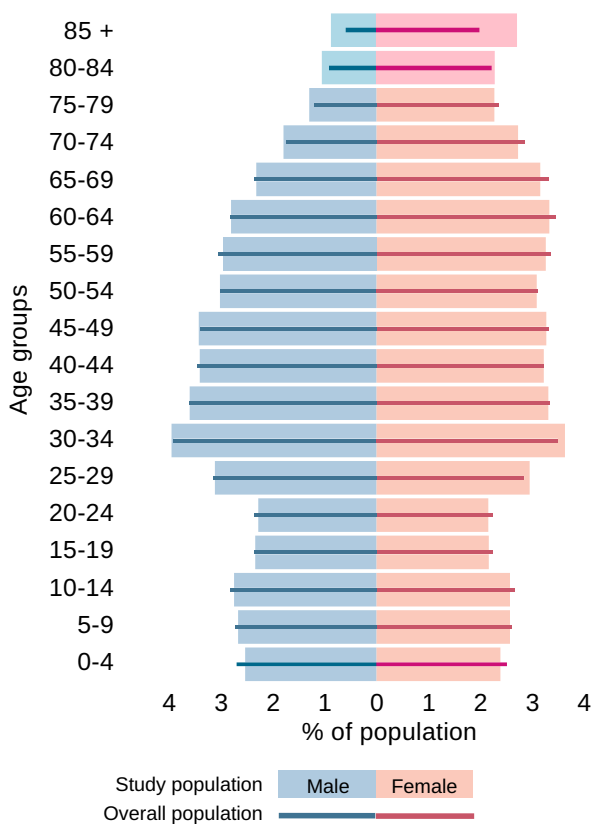


Figure 3. Population pyramids of an Estonian population in 2019 (lines) and the study sample (bars).

datasets enabled us to determine the year of birth for 530 patients who were missing this information in some of the source datasets, and thus, we were able to include them in the target database.

The distribution of source data across target tables is shown in Table 2. Most populated target tables are for measurements (32 230 620 entries) and conditions (20 351 014). The highest percentage of entries mapped to OMOP standardized vocabularies are for tables visit_occurrence and visit_detail (100%), condition_occurrence (99.9%), drug_exposure (98.7%), observation (97.1%), and measurement (90.8%).

Table 3 shows the mapping rate according to source vocabularies. Local service codes, ICD-10 codes, and LOINC codes were used the most in the data. For ICD-10 and LOINC codes, we covered almost all entries with standard concepts (100.0% and 99.3%, respectively). The coverage was also high for entries where local service codes were used (84.6%), although we only mapped 38.9% of source codes. 91% of drug product codes were mapped to RxNorm leaving 357 codes without a corresponding mapping. Among these 357 product codes, 202 distinct ATC codes were identified, and 143 of them were successfully mapped. Consequently, our mapping efforts resulted 98.7% mapping rate for drug entries.

The summary data of the output database can be found on a dedicated website.³³ The results of the DataQualityDashboard tool, which validates the plausibility, conformance, and completeness of the output dataset, are shown in Table S1. Out of the 3482 checks conducted, 3431 passed, and 51 failed. The failed checks were evaluated individually, and it was determined that their failure was expected. Five

plausibility checks failed because a gender-related clinical code was assigned to the wrong gender person. For example, 4 records with the concept of “primary malignant neoplasm of penis” were assigned to persons whose gender was female. For the completeness test, checks for 5 tables failed as the percentage of records where the standard concept value was “0” exceeded the threshold of 5%. This was due to the presence of unmapped records in the tables. Additionally, we had 41 checks failing due to the high proportion (>5%) of measurement values outside the range specified in DataQualityDashboard. Our investigation revealed that the values were within reasonable ranges (thus, no error in the data), and to our knowledge, the range windows have been removed in the next version of DataQualityDashboard.

Discussion

This article describes the integration process of claims data, EHRs, and prescriptions data into one complete patient-centered view. For a 10% random sample, 3 national health databases with complete records were linked and successfully transferred to OMOP CDM. Our experience shows that transferring health databases to OMOP CDM contains several challenges (Table 4). However, the outcome of the mapping and transformation process has a good quality and expands the possibilities for collaboration. To the best of our knowledge, this is one of the first papers of this kind and one of the largest by the proportion of the population of a country.

In total, we transformed over 100 million entries to standard concepts using standard OMOP vocabularies with the average mapping rate 95%. For conditions, observations, drugs, and measurements, the mapping rate was over 90%. In most cases, SNOMED CT was used as the target vocabulary. Similarly to others,^{11,13,16,21} we were not able to map all the records.

Previous studies have shown that one of the main difficulties during the transformation process is finding the relevant concepts.^{6,16,17,21,23} This is also in line with our experience. In case source vocabulary was already considered standard vocabulary according to OMOP CDM or mapping between source and standard vocabulary was provided, we used that vocabulary. However, it is crucial to stay aware of the changes in standard vocabularies and to be prepared to update mappings continuously. There is also a possibility that standard vocabulary becomes nonstandard over time. For instance, at the beginning of our mapping process, there were 2 standard vocabularies for cancer findings, but later, one of them was changed to nonstandard by the OHDSI community, which created additional work. It would be helpful if the OHDSI community could provide clear recommendations for determining the most suitable target vocabularies for mapping to avoid potential issues in the future. When mappings to standard vocabularies are not present, there are, unfortunately, no established guidelines for determining which vocabulary should be used. We recommend selecting the target vocabulary that the user is most familiar with. We agree with the previously reported assertions that the most problematic is the mapping of local code systems^{16,21,23} as ambiguity in source codes makes it difficult, if not impossible, to find appropriate target codes. This is the main reason why our mapping rate of local service codes is rather low (39%),

Table 2. Number of entries in OMOP CDM tables together with mapping rates.

OMOP CDM table	Total number of entries in table	The number of entries mapped to OMOP standard concepts	Mapping rate
Location	1	Not applicable	Not applicable
care_site	1820	Not applicable	Not applicable
person	149 364	Not applicable	Not applicable
death	8277	Not applicable	Not applicable
observation_period	149 364	Not applicable	Not applicable
visit_occurrence	18 281 120	18 281 120	100.0%
visit_details	48 002	48 002	100.0%
condition_occurrence	20 351 014	20 333 065	99.9%
procedure_occurrence	6 956 568	5 392 192	77.5%
drug_exposure	7 945 992	7 842 231	98.7%
device_exposure	77 842	47 296	60.8%
observation	15 203 064	14 762 978	97.1%
measurement	32 230 620	29 250 571	90.8%

Table 3. Mapped concepts and number of entries according to source vocabularies.

Source vocabulary	Number of source concepts	Number of mapped concepts	Mapped concepts (%)	Number of source entries	Number of mapped entries	Mapped entries (%)
ICD-10	9752	9751	100.0	22 738 540	22 738 530	100.0
LOINC	3169	2652	83.7	20 640 878	20 488 392	99.3
Drug product code	3946	3589	91.0	7 562 932	7 502 917	99.2
ATC ^a	202	143	70.8	60 015	12 682	21.1
Local service codes	2518	979	38.9	29 252 643	24 757 917	84.6
NCSP	3960	602	15.2	842 504	420 140	49.9
Other ^b	1396	1396	100.0	709 268	709 268	100.0

^a Only entries which were not mapped on drug product code level.

^b Cancer-related codes, pathology codes, and body measurements.

but it still covers the most frequently used local codes in Estonia.

In addition to incomplete mapping, information loss during the transformation process can also occur due to data structure.²³ Our experience showed that integrating data from various healthcare settings can result in overlap, where the same health event or episode may be represented in multiple datasets. In case of overlapping events, a decision had to be made as to whether to treat these multiple instances of the same event (eg, diagnosis) “as is” or to attempt to combine them into a single event in the OMOP CDM. After careful consideration, it was ultimately decided not to link the records and to use multiple instances of the event in the OMOP database. Several factors justified this decision. Firstly, the process of combining multiple records into a single event is complex and prone to error. Secondly, the transformation process would result in the loss of information about the exact source code and dataset, which may be necessary for quality control or specific studies. Finally, previous studies conducted using the OMOP CDM have not typically been concerned with the number of records or have required a minimum time interval between records, effectively addressing the issue of multiple close-time-range recordings of the same event.^{34,35}

Several actions can support and improve the transformation process. One of them is communication and collaboration with different experts and consortiums.³⁶ It is mandatory to consult with the medical personnel to understand the clinical practice and map the codes correctly. Still, in our case, despite our efforts and the inclusion of medical experts, more than half of the local codes remained

unmapped (61%). During the mapping process, we consulted with the OHDSI community about drug prescriptions data. According to the recommendations received, only drugs purchased from the pharmacy were mapped in our study. While information about prescribed medications may sometimes be more important than the fact of purchase, it was decided to exclude it to minimize confusion in the execution phase of future studies. In addition, participation in international projects and consortiums can provide insight into any additional data requirements, existing problems, and necessary mappings for specific studies. By having one modular codebase, this knowledge will accumulate and can be built upon in subsequent studies.

When planning a data transformation to OMOP CDM, it should be considered that transforming one or multiple linked datasets to a CDM cannot be taken as a one-time project but rather a continuous and iterative process requiring dedicated personnel, tools, and experience. This recommendation was previously highlighted by Candore et al³⁶ as well. For example, conducting a study on the created dataset can reveal issues in the data that were missed during the transformation. This means that some transformation steps must be repeated to improve the data quality. Due to the continuous and iterative nature of the transformation process, our experience highlights that it is essential to have the entire process as a version-controlled software code that can be reused, including any mappings created in previous studies. This will eventually lead to a gradual improvement in the quality of the dataset.

Despite our efforts, our work has some limitations that must be considered. Firstly, the described transformation pipeline with the created mappings is directly applicable to

Table 4. Main challenges and solutions of the current work.

Challenge	Example	Solution
The same health event is represented in several source datasets without a clear link between them, potentially leading to duplicates.	The same diagnosis code for a patient may be recorded in an EHR, claim, and prescription files. However, it may be difficult to link these documents to a single event due to the absence of a unique identifier for the case.	Transform each record as they are (even if duplicates) but add the provenance information to the record so one can use it when making cohorts.
No clear guidelines for choosing target vocabulary when multiple standard OMOP vocabularies are available. Additionally, there is no roadmap indicating which standard vocabularies may no longer be considered standard for OMOP CDM in the near future.	Physician Current Procedural Terminology Fourth Edition (CPT4) and SNOMED CT are both standard OMOP vocabularies for procedures; similarly, LOINC and SNOMED CT are for lab tests. The National Cancer Institute Thesaurus (NCIt) was a standard OMOP vocabulary at the beginning of our study, but not standard anymore.	Use the target vocabulary you are more familiar with. Keep in mind that what constitutes a standard OMOP vocabulary may change over time.
Hard to keep manual mapping files up to date as the standard target concepts change over time when updating the vocabularies.	Local code “9124,” which is used for vaccination against diphtheria and tetanus, was mapped to SNOMED CT code “73152006” (administration of diphtheria and tetanus vaccine). That target concept changed from standard to nonstandard at some point in time. Thus, we had to remap it to the concept code “1657590” from RxNorm vocabulary (diphtheria toxoid vaccine, inactivated/tetanus toxoid vaccine, inactivated injection).	Whenever updating the vocabulary, recheck the mappings in Usagi before running the transformation. Usagi automatically creates the list of nonstandard mappings so one can fix them before the actual data transformation.
Hard to keep track of all the historical coding versions of the same event to use similar target mapping for these.	The atypical squamous cells of undetermined significance (ASC-US) result of the Papanicolaou test have been recorded in our datasets by SNOMED CT code “39035006,” SNOMED morphology code “M-697102,” local codes “D,” “D1,” and “D1.1,” and in free-text format.	When working with historical codes and data, always check the most recent target code for this event to reuse the same code.
Broad source codes are hard to map to specific target codes.	Local code “7004” is used for all kinds of biopsies in the claims database. Also, “HPV test” (referring to human papillomavirus testing) or “eGFR” (estimated glomerular filtration rate) are noted in the text without further details on which particular test (LOINC code) was carried out.	Try to use additional information from the same medical record to specify the target code. For example, a diagnosis code referring to the prostate may help to map the biopsy to a more specific prostate biopsy.
Which of these events to transform—the prescription of the drug or the purchase? Or both?	After a drug is prescribed to the patient, they may or may not purchase it. Sometimes the buy-out happens several months after the prescription. Although the OMOP CDM allows recording both types of events separately, drug era calculation does not differentiate between these.	Prefer the purchase information as it better reflects what the patient may have consumed. Do not hesitate to consult with the OHDSI community and other research groups working with OMOP CDM, as their experience can give invaluable input on how to deal with your data most effectively.
ATC codes for drugs consisting of several ingredients map to nonstandard RxNorm codes.	ATC code “C09BX01” (perindopril, amlodipine, and indapamide; systemic) refers to an angiotensin-converting enzyme (ACE) inhibitor combination drug. It has 3 different ingredients. In OMOP CDM, there is no single standard concept for that combination drug, and mapping it into 3 separate ingredients (perindopril, amlodipine, indapamide) can lead to other problems.	Use extra information about the drug, such as product information, to find standard RxNorm codes for the mapping.
It is not systematically specified in the LOINC coding system what the expected results of a lab test are, making it difficult to decide on which LOINC answer code the results should be mapped to.	Depending on a particular lab test, the negative results can be given as “Negative,” “Not present,” “Not detected,” “Absent,” etc. For some tests, the expected results are not given at all in LOINC nomenclature.	In cases where the official result code is not specified in LOINC, use LOINC standard “negative” code.
It is difficult to achieve the best mapping quality without a complete understanding of the underlying medical practices.	Although we have a specific code for prostate biopsy in a local code system, it is rarely used. Broader “biopsy” is used instead.	Talk to medical personnel who can describe the underlying medical and data recording process.
Mapping is usually never 100% complete. In each study, there is a need to map some additional data.	For a cancer study, we need to extract tumor specific TNM and stage information from free text parts of the source data.	Build the whole mapping and transformation process as a repeatable software code and workflow so that each following study can reuse the mappings from previous studies. Be aware of the additional mapping needed and keep the necessary expertise in the team.

Estonian national datasets only. Secondly, the created dataset cannot be made publicly available. Also, as the dataset includes data from 2012 to 2019, the current observation period can be too short for some studies. At the same time, the age and gender distribution of the sample follow the whole population; thus, the dataset can be considered representative. To date, the work is still ongoing to improve the quality of the dataset and extract as much important information as possible. For example, the mapping coverage for NSCP classification of surgical procedures or device exposures (Tables 2 and 3) is currently modest, as these have not been the focus of our current research. Moreover, there is still a large amount of information stored as free text, which is being gradually extracted. In this research article, our focus has not been exhaustive in addressing the intricacies of data extraction from EHR documents. Given the depth and complexity of this topic, we think that this aspect warrants heightened consideration in future scientific literature.

As a result of the transformation process, we have created a rare example of Northern-Eastern European datasets mapped to OMOP CDM containing data from the 10% of Estonian population and almost all healthcare settings. To our knowledge, the harmonization and integration of these 3 national datasets have been a unique and innovative effort, even for a large OHDSI community. The usefulness of the dataset has been demonstrated through its application in various national and international studies and projects for generating evidence. For example, in Estonia, we have performed a study investigating the presence of HPV virus types and cervical cytology grades³⁷ and analyzed how artificial intelligence could be applied to health data for public service.³⁸ In addition, developing a tool for analyzing health event trajectories in any OMOP dataset³⁹ or participating in the study-a-thon of a project to harness big data in prostate cancer research⁴⁰ would not have been possible without the transformation process. We have validated the data linkage and the above describe repeatable approach in the PIONEER study, where the cohort of patients with newly diagnosed prostate cancer had an inclusion criterion requiring both a diagnosis and biopsy to be recorded.⁴¹ Using only EHR or prescription data would have yielded zero patients in Estonia while using only claims data would have yielded 235 patients. However, combining these 2 datasets resulted in a person count of 635 patients. The repeatable transformation scripts have been reused with only minor adjustments on independent new cohorts on the prescriptions of 110 000 asthma patients, insurance claims, and prescriptions of 400 000 COVID-19 patients with controls, and on all the medical data of more than 200 000 gene donors of Estonian Biobank. This all has significantly contributed to the efficient use of RWD.

Conclusion

This article contributes to the broader use of RWD. We have described our approach to link 3 central health databases in Estonia and successfully demonstrated transferring 10% of the data to OMOP CDM. The methods described can be applied to any future study using Estonian health data, and could potentially be used to convert the entire population's health data to OMOP CDM. Additionally, these principles can be applied beyond Estonia. Despite the challenges faced during the transformation process, our experience shows that OMOP CDM can be effectively used for healthcare data and

that the transformation can increase the opportunities for health data analysis and collaboration. Our work helps future researchers to transform linked databases into OMOP CDM more efficiently, ultimately leading to better RWE.

Acknowledgments

The data processing described in the article was carried out at the High Performance Computing Center of the University of Tartu.

Author contributions

M.O.: methodology, formal analysis, investigation, writing-original manuscript, writing-review & editing, software, and final approval for submission. S.T.: methodology, formal analysis, investigation, writing-review & editing, software, and final approval for submission. K.M. and M.P.: writing-original manuscript, writing-review & editing, and final approval for submission. H.-A.T., A.O., M.Lõ., J.H., M.H., H.S., and D.S.: methodology, writing-review & editing, software, and final approval for submission. M.La. and M.M.: methodology, writing-review & editing, and final approval for submission. J.V. and R.K.: writing-review & editing, final approval for submission, project administration, and funding acquisition. S.L.: methodology, writing-review & editing, final approval for submission and project administration. S. R.: writing-original manuscript, writing-review & editing, software, final approval for submission, project administration, and funding acquisition.

Supplementary material

[Supplementary material](#) is available at *JAMIA Open* online.

Funding

This work was supported by the Estonian Research Council (grant numbers PRG1844 and RITA1/02-96-11), the European Social Fund via IT Academy Program, and the European Regional Development Fund (Estonian Center of Excellence EXCITE, TK148). The European Health Data & Evidence Network has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) (grant agreement no. 806968). The JU receives support from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA.

Conflicts of interest

None declared.

Data availability

The data underlying this article cannot be shared publicly for the privacy of individuals that participated in the study. The data were obtained from national health databases and can be requested via Estonian Bioethics and Human Research Council (Eesti bioetika ja inimuurigute nõukogu).

References

- Hubbard T, Paradis R. *Real World Evidence: A New Era for Health Care Innovation*. NEHI—Network for Excellence in Health Innovation; 2015. Accessed January 14, 2023. https://www.nehi-us.org/writable/publication_files/file/rwe_issue_brief_final.pdf
- Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence—what is it and what can it tell us? *N Engl J Med*. 2016;375(23):2293-2297. <https://doi.org/10.1056/NEJMSb1609216>
- U.S. Food and Drug Administration, *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices. Guidance for Industry and Food and Drug Administration Staff*. 2017. Accessed January 14, 2023. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices>
- Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337. <https://doi.org/10.1016/j.jclinepi.2004.10.012>
- Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc*. 2015;22(3):553-564. <https://doi.org/10.1093/jamia/ocu023>
- Kent S, Burn E, Dawoud D, et al. Common problems, common data model solutions: evidence generation for health technology assessment. *PharmacoEconomics*. 2021;39(3):275-285. <https://doi.org/10.1007/s40273-020-00981-9>
- Blacketer C, Defalco FJ, Ryan PB, et al. Increasing trust in real-world evidence through evaluation of observational data quality. *J Am Med Inform Assoc*. 2021;28(10):2251-2257. <https://doi.org/10.1093/jamia/ocab132>
- Reinecke I, Zoch M, Reich C, et al. The usage of OHDSI OMOP—a scoping review. In: Röhrig R, Beißbarth T, König J, et al., eds. *Studies in Health Technology and Informatics*. IOS Press; 2021:95-103. <https://doi.org/10.3233/SHTI210546>
- OHDSI. *OHDSI Community Dashboard. Publication Analysis*. Accessed November 22, 2022. https://dash.ohdsi.org/publication_dashboard/
- OHDSI. *Our Journey. Where the OHDSI Community Has Been and Where We Are Going*. 2022.
- Papez V, Moinat M, Voss EA, et al. Transforming and evaluating the UK biobank to the OMOP common data model for COVID-19 research and beyond. *J Am Med Inform Assoc*. 2022;30(1):103-111. <https://doi.org/10.1093/jamia/ocac203>
- 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(1):13-22. <https://doi.org/10.1055/s-0039-3402754>
- Papez V, Moinat M, Payralbe S, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. *JAMIA Open*. 2021;4(3):ooab001. <https://doi.org/10.1093/jamiaopen/ooab001>
- Delanerolle G, Williams R, Stipancic A, et al. Methodological issues in using a common data model of COVID-19 vaccine uptake and important adverse events of interest: feasibility study of data and connectivity COVID-19 vaccines pharmacovigilance in the United Kingdom. *JMIR Form Res*. 2022;6(8):e37821. <https://doi.org/10.2196/37821>
- Maier C, Lang L, Storf H, et al. Towards implementation of OMOP in a German University Hospital Consortium. *Appl Clin Inform*. 2018;9(1):54-61. <https://doi.org/10.1055/s-0037-1617452>
- Kim J-W, Kim S, Ryu B, et al. Transforming electronic health record polysomnographic data into the Observational Medical Outcome Partnership's Common Data Model: a pilot feasibility study. *Sci Rep*. 2021;11(1):7013. <https://doi.org/10.1038/s41598-021-86564-w>
- Paris N, Lamer A, Parrot A. Transformation and evaluation of the MIMIC database in the OMOP common data model: development and usability study. *JMIR Med Inform*. 2021;9(12):e30970. <https://doi.org/10.2196/30970>
- Ji H, Kim S, Yi S, et al. Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. *J Biomed Inform*. 2020;107:103459. <https://doi.org/10.1016/j.jbi.2020.103459>
- Sathappan SMK, Jeon YS, Dang TK, et al. Transformation of electronic health records and questionnaire data to OMOP CDM: a feasibility study using SG_T2DM dataset. *Appl Clin Inform*. 2021;12(4):757-767. <https://doi.org/10.1055/s-0041-1732301>
- Haberson A, Rinner C, Schöberl A, et al. Feasibility of mapping Austrian health claims data to the OMOP common data model. *J Med Syst*. 2019;43(10):314. <https://doi.org/10.1007/s10916-019-1436-9>
- Biedermann P, Ong R, Davydov A, et al. Standardizing registry data to the OMOP common data model: experience from three pulmonary hypertension databases. *BMC Med Res Methodol*. 2021;21(1):238. <https://doi.org/10.1186/s12874-021-01434-3>
- Yu Y, Jiang G, Brandt E, et al. Integrating real-world data to assess cardiac ablation device outcomes in a multicenter study using the OMOP common data model for regulatory decisions: implementation and evaluation. *JAMIA Open*. 2023;6(1):ooac108. <https://doi.org/10.1093/jamiaopen/ooac108>
- Zhou M, Fukuoka Y, Goldberg K, et al. Applying machine learning to predict future adherence to physical activity programs. *BMC Med Inform Decis Mak*. 2019;19(1):169. <https://doi.org/10.1186/s12911-019-0890-0>
- EHDEN. 2022—European Health Data Evidence Network. Accessed January 14, 2023. <https://www.ehden.eu/>
- EHDEN DPC. *A Federated Network of Data Partners*. Accessed January 14, 2023. <https://www.ehden.eu/datapartners/>
- Estonian Health Insurance Fund. *Estonian Health Insurance Fund Annual Report of Financial Year 2020*. Estonian Health Insurance Fund; 2021.
- Kõnd K, Lilleväli A. E-prescription success in Estonia: the journey from paper to pharmaco-genomics. *Eurohealth*. 2019;25(2):18-20.
- National Library of Medicine. *RxNorm*. Accessed January 14, 2023. <https://www.nlm.nih.gov/research/umls/rxnorm/index.html>
- OHDSI. *ATHENA OHDSI Vocabularies Repository*. Accessed January 14, 2023. <http://athena.ohdsi.org>
- Schuemie M. *Usagi*. Accessed January 14, 2023. <https://github.com/OHDSI/Usagi>
- Automated Characterization of Health Information at Large-Scale Longitudinal Evidence Systems (ACHILLES)*. Accessed January 14, 2023. <https://github.com/OHDSI/Achilles>
- CdmInspection. *R Package to Support Quality Control Inspection of An OMOP-CDM Instance*. Accessed January 14, 2023. <https://github.com/EHDEN/CdmInspection>
- Dashboard of Estonian MAITT Dataset*. Accessed January 14, 2023. <http://omop-apps.cloud.ut.ee/ShinyApps/MAITTOverview/#/MAITT/dashboard>
- Omar MI, Roobol MJ, Ribal MJ, et al.; PIONEER Consortium. Introducing PIONEER: a project to harness big data in prostate cancer research. *Nat Rev Urol*. 2020;17(6):351-362. <https://doi.org/10.1038/s41585-020-0324-x>
- Yang C, Williams RD, Swerdel JN, et al. Development and external validation of prediction models for adverse health outcomes in rheumatoid arthritis: a multinational real-world cohort analysis. *Semin Arthritis Rheum*. 2022;56:152050. <https://doi.org/10.1016/j.semarthrit.2022.152050>
- Candore G, Hedenmalm K, Slattery J, et al. Can we rely on results from IQVIA medical research data UK converted to the Observational Medical Outcome Partnership Common Data Model?: a

- validation study based on prescribing codeine in children. *Clin Pharmacol Ther.* 2020;107(4):915-925. <https://doi.org/10.1002/cpt.1785>
37. Uusküla A, Oja M, Tisler A, et al. Prevaccination prevalence of type-specific human papillomavirus infection by grade of cervical cytology in Estonia. *JAMA Network Open.* 2023;6(2):e2254075. <https://doi.org/10.1001/jamanetworkopen.2022.54075>
 38. Solvak M, Vilo J, Reisberg S, et al. RITA MAITT—Programmi RITA tegevuse 1 projekti „Masinõppe ja AI toega teenused “lõpparuanne” [Final Report of the Project “Machine Learning and AI Supported Services” of RITA Programme Task 1]. 2022. Accessed January 14, 2023. <https://www.etis.ee/Portal/Publications/Display/f077ca2a-db43-41e9-8076-51e3e2fb7ac3>
 39. Künnapuu K, Ioannou S, Ligi K, et al. Trajectories: a framework for detecting temporal clinical event sequences from health data standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *JAMIA Open.* 2022;5(1): ooac021. <https://doi.org/10.1093/jamiaopen/ooac021>
 40. Beyer K, Moris L, Lardas M, et al.; PIONEER Consortium. Updating and integrating core outcome sets for localised, locally advanced, metastatic, and nonmetastatic castration-resistant prostate cancer: an update from the PIONEER consortium. *Eur Urol.* 2022;81(5):503-514. <https://doi.org/10.1016/j.eururo.2022.01.042>
 41. Gandaglia G, Pellegrino F, Golozar A, et al.; PIONEER Consortium. Clinical characterization of patients diagnosed with prostate cancer and undergoing conservative management: a PIONEER analysis based on big data. *Eur Urol.* 2023. <https://doi.org/10.1016/j.eururo.2023.06.012>