Enhancing Local Vocabulary into OMOP Vocabulary based on the Semi-Automated Framework: Korean EDI Case Study

Yiju Park[1,2], Jinwoo Yoon[1,2], Seojeong Shin[2], Oleg Zhuk[3], Anna Ostropolets[4,5], Seng Chan You[1,2]

[1]Department of Biomedical Systems Informatics, Yonsei University, Seoul, Korea
[2]Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Korea
[3]Odysseus, an EPAM Company, Cambridge, USA
[4]Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA
[5]Department of Biomedical Informatics, Columbia University, NY, USA

Background

In Korea, the Electronic Data Interchange (EDI) code system, managed by the Health Insurance Review and Assessment Service (HIRA), is widely used for insurance claims under the fee-for-service system. The EDI codes represent drugs, devices, medical services, and have become the *de facto* standard in Korean electronic medical record (EMR) systems. However, the EDI vocabulary has two major limitations as a controlled vocabulary system: 1) Lack of concept permanence: The EDI lacks unique identifiers, allowing expired codes to be reassigned to new, completely different concepts, which compromises the consistency of data interpretation over time. 2) Semantic inconsistencies: In the EDI system, the same code can be used with different meanings and can disrupt data integration and interpretation between domains.

To address these challenges and improve the interoperability of EDI code system, Seong et al.[1] developed a process to integrate the EDI vocabulary into the OMOP vocabulary. In this previous work, components of controlled vocabulary including domain, relation, and unique non-sematic concept identifiers were formulated for EDI vocabulary to be integrated into OHDSI vocabulary. While this study made significant progress, it had limitations. First, it only utilized EDI data from October 2019, which limited the coverage and reproducibility of the transformed vocabulary. Second, the mapping of EDI concepts to Standard Concepts of OMOP vocabulary was insufficient. Lastly, there was no established vocabulary quality check protocol, making it challenging to validate the quality of the generated concepts.

This study aims to improve comprehensiveness of the previous work, integrating Korean EDI codes to the OMOP vocabulary, covering the entire EDI concepts spanning from November 2000 to May 2024. We also focus on implementing persistent identifiers, and mapping EDI concepts to standard concepts, which is essential for ensuring interoperability within the OMOP vocabulary ecosystem. We also leverage the OHDSI Vocabulary Quality Assurance and Control (QA/QC) protocol to validate the quality of the generated concepts.

Methods

This study collected all EDI code data published through the HIRA website[2] up to May 2024, significantly expanding the EDI vocabulary coverage compared to the previous study[1]. For instance, while the previous research only used EDI data from a single month (October 2019), this study collected EDI codes spanning nearly 24 years, from November 2000 to May 2024. This extensive data collection allowed us to capture the changes and deletions of codes in the EDI vocabulary over time, providing a more comprehensive and representative dataset for conversion to the OMOP vocabulary. To facilitate this process, we enhanced a package called 'SYNC', which is designed to handle large-scale, longitudinal data. The package has been made publicly available through GitHub[6] to ensure transparency and reproducibility of the research results.

The semi-automated 'SYNC' process involves four main steps: 1) Scraping data from HIRA by crawling the website, 2) Yielding (or labeling) data set into relevant domains: Device, Drug, Procedure, and Measurement, 3) Nesting the domains into a hierarchical structure (for example, we assigned the Procedure/Proc Hierarchy concept class based on the 5-digit numbers of EDI codes, referring to the Procedure domain classification), 4) Converting to English using Google translation API.

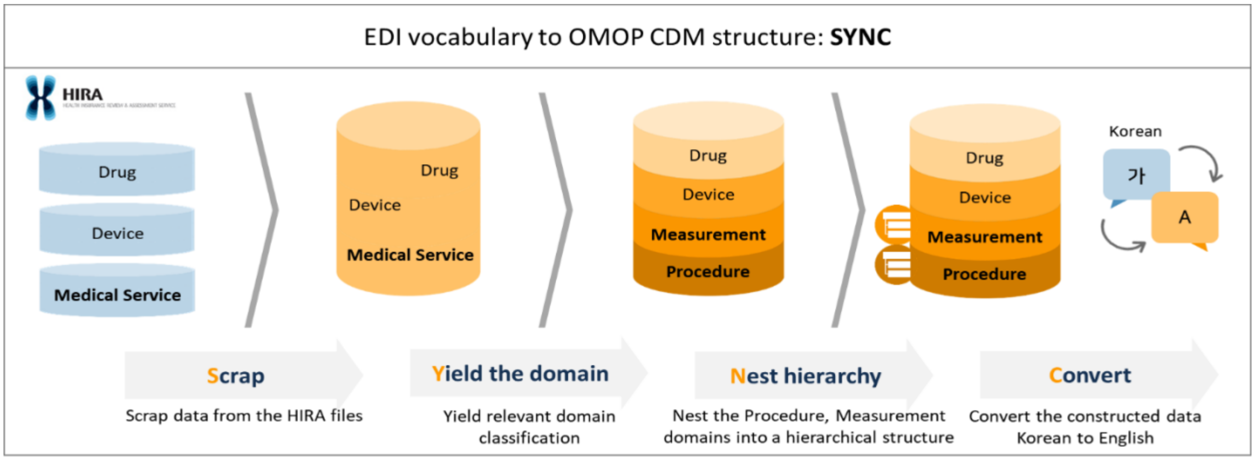
EDI vocabulary to OMOP CDM structure: SYNC
HIRA
Drug
Device
Medical Service
Drug
Device
Medical Service
Drug
Device
Measurement
Procedure
Drug
Device
Measurement
Procedure
Korean
가
A
Scrap
Scrap data from the HIRA files
Yield the domain
Yield relevant domain classification
Nest hierarchy
Nest the Procedure, Measurement domains into a hierarchical structure
Convert
Convert the constructed data Korean to English

Figure 1. The process of the semi-automated package 'SYNC'.

Building upon the previous study, we improved processes such as crawling the HIRA website to collect large-scale data, prioritizing the assignment of codes to Procedure and Measurement over the Device domain to resolve code duplication.

After applying the SYNC framework, EDI concepts were mapped to standard concepts following OHDSI-Korea community guidelines[3]. The mapping process involved manual mapping and review by medical informatics experts to ensure accuracy and consistency.

The mapped EDI vocabulary was then incorporated into the OMOP Vocabularies through the Community Contribution Process[4]. This process involves rigorous Quality Assurance and Control (QA/QC) checks on various aspects of the vocabulary, including its structure, content, and mapping to standard concepts. We cooperated with the OHDSI Vocabulary Team to address feedback, refine mappings, and correct errors identified during the QA/QC process. Once all quality checks were passed, the vocabulary is published in the OHDSI Vocabulary Repository, making it accessible in Athena.

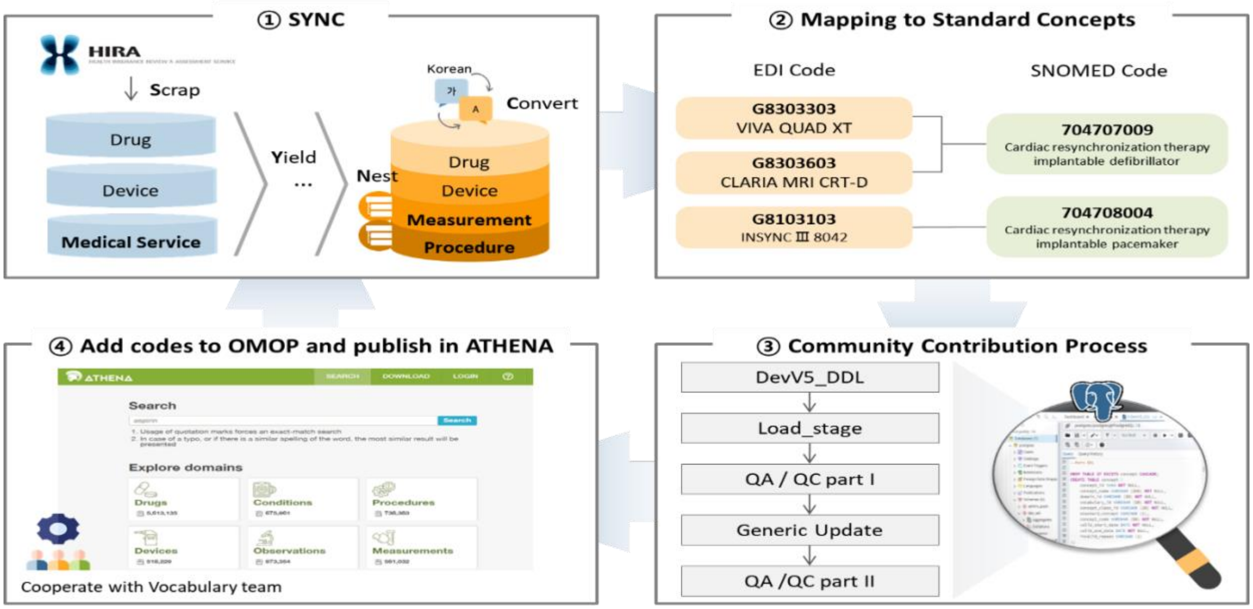

Figure 2. The overall process of this study.

Results

A total of 620,642 EDI codes uploaded to the HIRA website from November 2000 to May 2024 were collected using SYNC process. The distribution of EDI codes across domains was as follows: Procedure (71.5%), Measurement (10.6%), Drug (10.6%), and Device (7.3%).

Of the collected EDI codes, 377,798, (60.9%) were mapped to Standard Concepts. The mapping coverage varied across domains, with the highest coverage in the Drug domain (94.0%), followed by Device (80.0%), Procedure (62.6%), and Measurement (2.5%). As shown in Figure 2, the mapping coverage by domain is compared with the results from the previous study.

	Seong et al.(2021)			This study		
	Source code	Source Mapped	Mapped Percent(%)	Source code	Source Mapped	Mapped Percent(%)
Drug	23,231	0	0.00	65,981	62,038	94.02
Device	19,813	0	0.00	45,131	36,114	80.02
Procedure	249,785	37,869	15.16	444,021	277,982	62.61
Measurement	20,602	675	3.28	65,508	1,664	2.54
Total	313,431	38,544	12.30	620,641	377,798	60.87

Table 1. Results of Domain-specific Standard Concept Mapping: A Comparison with Previous Study.

The finalized OMOP CDM-structured EDI vocabulary was shared with the OHDSI Vocabulary team and committed to the OHDSI GitHub repository[5] for its reproducibility and transparency. Additionally, we upload the SYNC package to GitHub[6] to share our methodology.

Conclusion

Through this study, we have significantly improved the process of integrating EDI codes into the OMOP CDM. By enhancing the SYNC process, we have expanded the temporal range of data and substantially increased the data volume compared to previous research. Notably, we have improved the proportion of codes mapped to OMOP standard concepts from 12.3% to 60.9% and refined the code assignment process to address inter-domain duplication issues. This research has enhanced the international interoperability of Korean healthcare data. Furthermore, to address the lack of concept permanence and semantic inconsistencies in the EDI system, we plan to incorporate the latest EDI codes, which are continuously promulgated, into the OHDSI controlled vocabulary.

However, this study has some limitations. First, not all concepts could be mapped to standard concepts due to the large volume of EDI codes. As a result, we were unable to load all the collected EDI data into ATHENA. Second, the EDI vocabulary allows identifier duplication across different domains, while the OMOP CDM requires unique codes. To address this, we prioritized Procedure and Measurement domains during the loading process to ensure source identifier uniqueness, but this prioritization may have led to some information loss in Device domain.

Future research should focus on further improving mapping coverage, exploring strategies for resolving code duplication issues in the source vocabulary without missing any codes.

References

1. Seong Y, You SC, Ostropolets A, Rho Y, Park J, Cho J, Dymshyts D, Reich CG, Heo Y, Park RW. Incorporation of Korean electronic data interchange vocabulary into observational medical outcomes partnership vocabulary. Healthcare Informatics Research. 2021 Jan;27(1):29.
2. Health Insurance Review and Assessment Service website [Internet]. Available: https://www.hira.or.kr/main.do
3. OHDSI-Korea Vocabulary guidelines, GitHub. [Internet]. Available: https://github.com/ohdsi-korea/OmopVocabularyKorea
4. OHDSI Vocabulary Community Contribution Process, GitHub. [Internet]. Available: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Community-contribution
5. Codes for integrating EDI codes into ATHENA for OMOP Vocabulary, GitHub. [Internet] Available: https://github.com/OHDSI/Vocabulary-v5.0/tree/EDI_aug_2024/EDI
6. Semi-automated EDI package, ‘SYNC’. GitHub [Internet]. Available: https://github.com/dr-you-group/SYNC_EDItoOmopPackage