Original Article

Healthc Inform Res. 2021 January;27(1):29-38.
https://doi.org/10.4258/hir.2021.27.1.29
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Incorporation of Korean Electronic Data Interchange Vocabulary into Observational Medical Outcomes Partnership Vocabulary

Yeonchan Seong[1,2,*], Seng Chan You[1,*], Anna Ostropolets[3], Yeunsook Rho[4], Jimyung Park[5], Jaehyeong Cho[5], Dmitry Dymshyts[6], Christian G. Reich[7], Yunjung Heo[8], Rae Woong Park[1,5]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea
[2]Department of Sociology, Yonsei University, Seoul, Korea
[3]Department of Biomedical Informatics, Columbia University, New York, NY, USA
[4]Health Insurance Review & Assessment Service, Wonju, Korea
[5]Deparment of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea
[6]Odysseus Data Services Inc., Cambridge, MA, USA
[7]Real Wolrd Solutions, IQVIA, Cambridge, MA, USA
[8]Department of Medical Humanities and Social Medicine, Ajou University School of Medicine, Suwon, Korea

Objectives: We incorporated the Korean Electronic Data Interchange (EDI) vocabulary into Observational Medical Outcomes Partnership (OMOP) vocabulary using a semi-automated process. The goal of this study was to improve the Korean EDI as a standard medical ontology in Korea. **Methods:** We incorporated the EDI vocabulary into OMOP vocabulary through four main steps. First, we improved the current classification of EDI domains and separated medical services into procedures and measurements. Second, each EDI concept was assigned a unique identifier and validity dates. Third, we built a vertical hierarchy between EDI concepts, fully describing child concepts through relationships and attributes and linking them to parent terms. Finally, we added an English definition for each EDI concept. We translated the Korean definitions of EDI concepts using Google.Cloud.Translation.V3, using a client library and manual translation. We evaluated the EDI using 11 auditing criteria for controlled vocabularies. **Results:** We incorporated 313,431 concepts from the EDI to the OMOP Standardized Vocabularies. For 10 of the 11 auditing criteria, EDI showed a better quality index within the OMOP vocabulary than in the original EDI vocabulary. **Conclusions:** The incorporation of the EDI vocabulary into the OMOP Standardized Vocabularies allows better standardization to facilitate network research. Our research provides a promising model for mapping Korean medical information into a global standard terminology system, although a comprehensive mapping of official vocabulary remains to be done in the future.

Keywords: Medical Informatics, Controlled Vocabulary, National Health Programs, Biological Ontologies, Knowledge Bases

Submitted: November 4, 2020, **Revised:** 1st, January 4, 2021; 2nd, January 23, 2021, **Accepted:** January 23, 2021

Corresponding Author
Yunjung Heo
Department of Medical Humanities and Social Medicine, Ajou University School of Medicine, 164 World cup-ro, Yeongtong-gu, Suwon 16499, Korea. Tel: +82-31-219-5285, E-mail: mellisa7@aumc.ac.kr (https://orcid.org/0000-0001-5708-1428)
*These authors contributed equally to this work.

I. Introduction

A standardized and controlled vocabulary in a national healthcare system facilitates semantic interoperability and collaborative research [1]. For medical diagnosis, the Korean Standard Classification of Diseases and Causes of Death (KCD-7), an extension of the tenth revision of the International Statistical Classification of Diseases and Related Health Problems 10th revision (ICD-10), is widely acknowledged as the *de facto* standard vocabulary because it is a mandatory terminology for claims operations. However, there has been no widely accepted standardized vocabulary system that incorporates drugs, medical services, and devices in Korea. The Korean Standard Terminology of Medicine (KOSTOM) was developed in 2004 to provide a standardized and comprehensive vocabulary of medical terminology [2]. However, because of a lack of commitment and inadequate publicity, the KOSTOM vocabulary has been seldom adopted in routine clinical practice or in big data analytics in medicine and healthcare [3].

The Health Insurance Review and Assessment Service (HIRA) has developed and maintains the Electronic Data Interchange (EDI) code system, or EDI vocabulary, to classify and identify drugs, medical services, and devices. HIRA mandates use of this vocabulary to obtain reimbursement in the fee-for-service system. For this reason, every Korean Electronic Health Record (EHR) system uses the EDI vocabulary for most drugs, medical procedures, and devices. However, most hospitals have developed their own medical vocabulary systems because of the limited granularity of the EDI vocabulary [4]. Furthermore, the EDI vocabulary has not been acknowledged as a standard vocabulary in the way that the Current Procedural Terminology, fourth edition has in the United States because the quality of the EDI has never been audited. To standardize this *de facto* Korean medical vocabulary, there was an effort to map the EDI vocabulary to the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) [5]. Nonetheless, this did not lead to substantive quality improvement of the EDI vocabulary itself.

1. Challenges in EDI Vocabulary as a Controlled Vocabulary

We identified the following five main problems disrupting the EDI's maintenance as a controlled medical vocabulary: lack of concept identifier (ID) version control, lack of ID permanence, use of semantic concept identifiers, non-unique identifiers, and lack of formal definitions.

First, the EDI has no controlled life cycle for its terms. The validity dates for EDI codes are not recorded in the official monthly announcements, but newly added and expired codes are announced in monthly announcements. Second, the identifiers and concepts of the EDI are not permanent. There are EDI vocabularies that are no longer used because of having expired or having been replaced by other vocabularies. We have confirmed that some of their expired codes have been reused in other vocabularies. Outdated EDI identifiers can be assigned to new concepts. That is, outdated EDI IDs can be assigned to new concepts. Third, the EDI vocabulary uses semantic concept identifiers. For example, the EDI ID of a drug includes information on the country, company, unit, and packaging type. This ontological system makes it difficult to apply a single rule if the number of tracked contents exceeds the digits allotted to represent the specific contents. Fourth, the EDI vocabulary has some duplicated identifiers because there is no unified EDI encoding system across domains. For example, 13 codes are duplicated between medical services and devices. Among these, "Chest [Direct], radiologist reading" in medical services and "TRIMO" in devices share the EDI ID G2101006. Fifth, although the EDI includes a modifier for reimbursing the additional price of service (e.g., emergency services or nighttime services) according to the national reimbursement policy, the concept definitions do not include information related to the modifiers. For example, the EDI ID N0333 means "Craniotomy or Craniectomy for Decompression." If the identical medical service is performed at night, it is recorded as EDI ID N0333010, but the conceptual definition remains "Craniotomy or Craniectomy for Decompression." Furthermore, Korean definitions of items in the EDI vocabulary vary across time, usually because of non-semantic punctuation.

2. Observational Medical Outcomes Partnership Vocabulary

Observational Health Data Sciences and Informatics (OHDSI) is an international, multi-stakeholder, interdisciplinary initiative for collaborative medical research, which uses an open-source standardized data structure and provides analytic solutions. As a successor to the Observational Medical Outcomes Partnership (OMOP), OHDSI adopts the OMOP common data model (CDM) as its standard data structure and the OMOP vocabulary as its standard semantics [6]. Multiple medical vocabulary systems are organized in the united controlled vocabulary system of the OMOP-CDM to provide comprehensive coverage for diverse healthcare databases across countries [7]. The OMOP vocabulary system comprises standard and non-standard vocabularies across various healthcare data domains, including condition (a medical diagnosis), drug, procedure, measurement, and device. For the condition domain, the SNOMED-CT and ICD-O (International Classification of Diseases for Oncology) vocabularies are used for the standard vocabulary, and ICD-10, ICD-10-CM, or KCD7 are classified as non-standard vocabulary. The OHDSI vocabulary subgroup evolved and maintained both standard and non-standard OMOP vocabulary based on desiderata for controlled medical vocabularies, such as concept orientation, concept permanence, non-semantic concept identifiers, polyhierarchy, formal definitions, multiple granularities, and graceful evolution [8].

3. Objectives

Our ultimate goal was to improve the EDI vocabulary for a controlled and standardized vocabulary system. For this purpose, we incorporated the EDI vocabulary into the OMOP Standardized Vocabulary through a semi-automated process.

II. Methods

For this study, we used the EDI concept list that was released on the HIRA website in October 2019. The EDI has separate vocabularies for drugs, medical services, and devices. These three domains have no unified system in the EDI vocabulary. A complete list of valid EDI codes in each of these three domains is independently released with a description every month. Figure 1 presents the overall process. First, we assigned a permanent, non-semantic, and unique concept identifier to each EDI concept. A "permanent" identifier refers to a concept identifier that will not be re-assigned to a new concept, and the identifier will contain expired data after the concept expires. A "non-semantic" and "unique" identifier means that the concept identifier *per se* is a random unique number without any meaningful information. Second, we established correspondences for all EDI vocabulary items for the four domains of the OMOP (drug, procedure, measurement, and device) with a hierarchy. Third, we translated the Korean definitions of EDI terms into English by leveraging Google Cloud Translation API to generate formal English definitions of all concepts.

We built a semi-automated process to incorporate the EDI vocabulary into the OMOP Standardized Vocabulary, including code cleaning, classification, building hierarchy, and vocabulary insertion in the OMOP-CDM version 5.3.1 database. We deployed the open-source click-to-run R software, EdiToOmop, found on the OHDSI's official GitHub repository [9].

1. Classification of Domains, Application of Management Systems and Building Hierarchy

Clinical events are classified into the domains of drug, device, condition, and procedure in OMOP. EDI concepts are divided into drugs, devices, and medical services, but the scope of medical services is too broad for the OMOP Standardized Vocabularies. Because of this discrepancy in domain classification between the EDI and OMOP Standardized Vocabularies, we subclassified EDI medical services into procedures and measurements to match the OMOP domains. To ensure that each concept's meaning would be clear and unique, we added more descriptive matter to the concept definitions to explain the modifier codes of the original EDI ID, such as emergency use.

Once registered in the OMOP Standardized Vocabularies, a permanent, unique, and non-semantic numeric OMOP identifier was assigned to each EDI concept. This identifier, called a concept ID, prevented duplication and tracked the concept's history from the first appearance to the depreca-

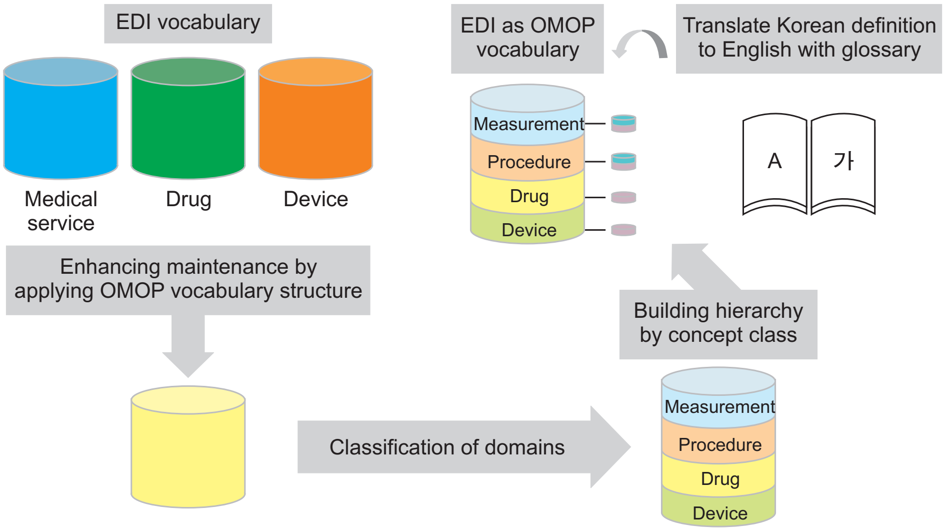

EDI vocabulary
Medical service
Drug
Device
Enhancing maintenance by applying OMOP vocabulary structure
Classification of domains
Measurement
Procedure
Drug
Device
Building hierarchy by concept class
EDI as OMOP vocabulary
Measurement
Procedure
Drug
Device
Translate Korean definition to English with glossary
A
가

Figure 1. The overall process. After incorporating HIRA's EDI vocabulary into the OMOP vocabulary, the domains of the concepts were classified. The hierarchical structures and English definitions were then added. EDI: Electronic Data Interchange, OMOP: Observational Medical Outcomes Partnership.

tion of EDI concepts. Three attributes define the validity of concepts in the OMOP Standardized Vocabularies: "valid start date," "valid end date," and "invalid reason." When an EDI concept is newly registered or deprecated, the term's date is updated or expired and is recorded. If a concept is valid, the "invalid reason" for the concept is recorded as "NULL." If a concept is replaced by another concept or deleted, the "invalid reason" for the concept is recorded as "U" or "D," respectively.

The OMOP Standardized Vocabulary provides vertical and horizontal hierarchical relationships between concepts. In this project, we built a formal vertical hierarchy for EDI concepts. As with the ICD-9 and ICD-10 code system, the first five digits of the EDI IDs in the medical service domain represent the ancestor terms for longer, descendent EDI IDs. The remaining digits are usually added as modifiers to the same service for reimbursement. Thus, the descent concept contains all of the information for the ancestor concept, creating a vertical hierarchy.

2. Translation

For incorporation into the OMOP Standardized Vocabularies, the English definition for each EDI term is essential. We identified 266,140 concept definitions without an English description in the EDI vocabulary domains of medical services and devices. The translation of these terms involved three steps. To increase efficiency, we leveraged a Google translation tool. We used the Google.Cloud.Translation.V3, a .NET client library in the Google Cloud Translation API for the initial translation. Because Google-translated definitions may have misrepresented the meaning of a Korean term or may not have recognized an abbreviated term, two registered nurses reviewed and modified the English definitions. As a second modification, we developed a glossary for Korean words that were often not translated correctly into English by the software. Google Translation API provides customized translation functions that refer to a glossary. We created a glossary containing 749 terms of devices and 6,079 terms of service. This includes modifiers for reimbursing the additional price of service. Referring to the glossary, a secondary translation was conducted for 266,140 words that needed to be retranslated. After the secondary translation using the glossary, a medical worker audited the translation to ensure precision.

3. Auditing of Vocabulary

Qualitative criteria indicate that our EDI vocabulary restructuring process improved data quality for the health terminology system. Cimino [8], Chute et al. [10], and Rosenbloom et al. [11] presented qualitative evaluation criteria for terminology. Additionally, Lee [12] synthesized the criteria and included an index to determine whether the terminology system could support multiple languages. Based on Lee's study [12], we defined the following 11 criteria for evaluating terminology and evaluating the incorporation of the EDI vocabulary into the OMOP Standardized Vocabularies: concept orientation, concept permanence, coverage, relation, multiple hierarchy, compositionality, non-semantic concept identifiers, version control, formal definitions, synonyms uniquely identified and mapped to relevant concepts, and multi-language.

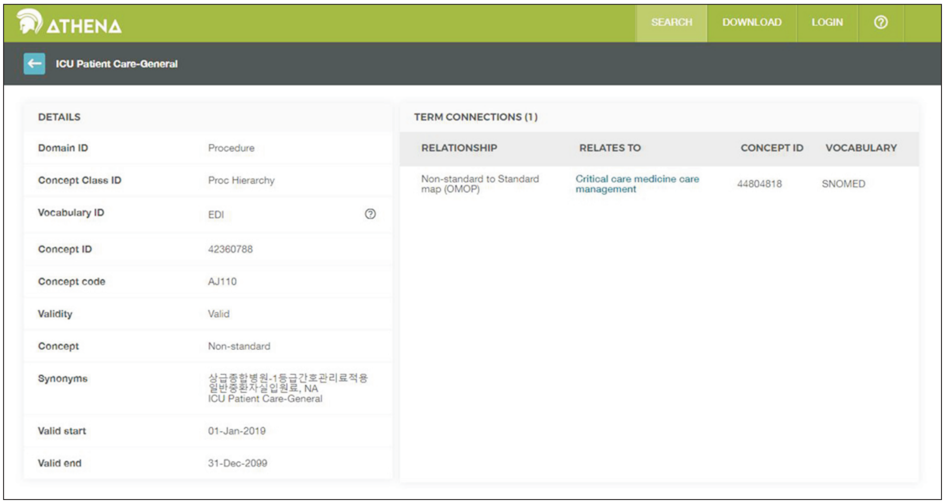
Another aspect of the EDI in the OMOP Standardized Vocabularies is the hierarchical relationships that we constructed. Furthermore, a mapping relation from non-standard to standard has been built. Thus, EDI concepts acquire relationships with other standard vocabularies. For example, the concept "ICU Patient Care-General" (OMOP Concept ID: 42360788) in the EDI is related to the concept of "Critical Care Medicine Care Management" (OMOP Concept ID: 44804818) in SNOMED-CT as shown in Figure 2.

The criterion for formal definition is related to multiple hierarchies. In the converted EDI vocabulary, each term acquires a formal definition, allowing concepts to have relationships with other concepts. For example, hierarchy defines parent/child relationships between concepts, such that "Intravenous Catheterization for Hemodialysis" (EDI ID: O7016) is the parent concept for "Intravenous Catheterization for Hemodialysis, second surgery" (EDI ID: O7016001).

A given unique integer identifier managed synonyms for unique concepts, and related concepts were mapped to each other. Moreover, we have given EDI terms of unique English versions. Through the EdiToOmop package, newly added or deprecated EDI IDs can be updated in the OMOP Standardized Vocabularies semi-automatically.

III. Results

The R package EdiToOmop was developed to automate the incorporation of the EDI vocabulary into the OMOP Standardized Vocabularies. Of 313,453 EDI concepts, 313,431 were incorporated, with 270,387 medical services classified as measurements or procedures. Of the 12,991 measurement codes, 1,301 were classified as ancestor codes, and 11,681 were classified as descent codes. For procedure codes, of 257,396 concepts, 7,038 were classified as ancestor codes, and 250,358 were classified as descent codes. Table 1 pres-



ICU Patient Care-General

DETAILS

Domain ID	Procedure
Concept Class ID	Proc Hierarchy
Vocabulary ID	EDI
Concept ID	42360788
Concept code	AJ110
Validity	Valid
Concept	Non-standard
Synonyms	상급종합병원-1등급간호관리료적용 일반중환자실입원료, NA ICU Patient Care-General
Valid start	01-Jan-2019
Valid end	31-Dec-2099

TERM CONNECTIONS (1)

RELATIONSHIP	RELATES TO	CONCEPT ID	VOCABULARY
Non-standard to Standard map (OMOP)	Critical care medicine care management	44804818	SNOMED

Figure 2. The concept "ICU Patient Care-General" (OMOP Concept ID: 42360788) in the EDI is related to the concept of "Critical Care Medicine Care Management" (OMOP Concept ID: 44804818) in SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms). ICU: intensive care unit, OMOP: Observational Medical Outcomes Partnership.

Table 1. Number of reclassified Korean EDI vocabulary concepts

EDI		OMOP		Concept classification	
Domain	Number of concepts (%)	Domain	Number of concepts (%)	Concept ID	Number of concepts (%)
Medical service	270,387 (86.3)	Measurement	12,991 (4.1)	Measurement	11,681 (3.7)
				Meas. class	1,310 (0.4)
		Procedure	257,396 (82.1)	Procedure	250,358 (79.9)
				Proc. hierarchy	7,038 (2.2)
Drug	23,231 (7.4)	Drug	23,231 (7.4)	Drug product	23,231 (7.4)
Device	19,813 (6.3)	Device	19,813 (6.3)	Device	19,813 (6.3)
Total	313,431 (100)	-	313,431 (100)	-	313,431 (100)

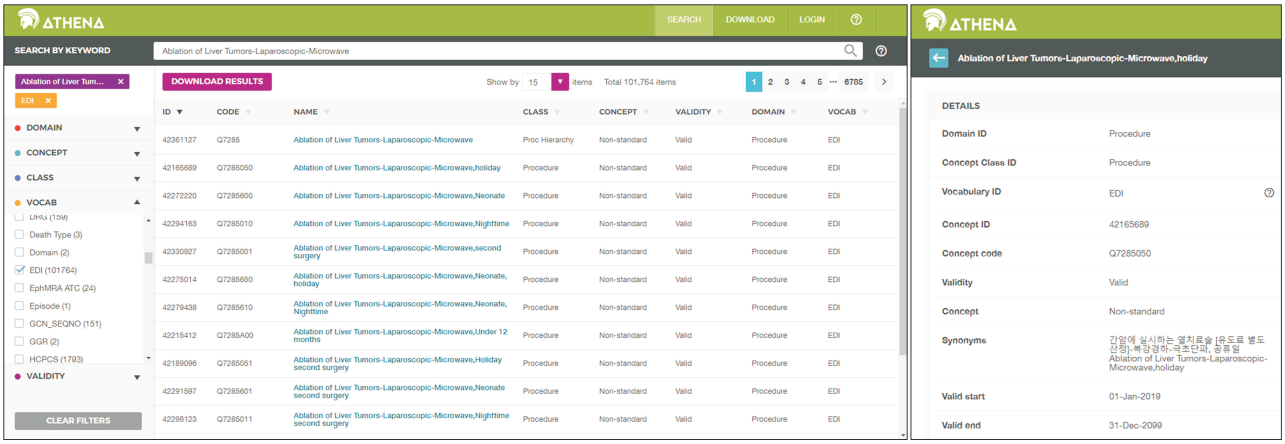
EDI: Electronic Data Interchange, OMOP: Observational Medical Outcomes Partnership.

ents the numbers of concepts in the original EDI vocabulary and the reclassified domains using a simple hierarchy for incorporation into the OMOP Standardized Vocabularies.

Redacted EDI concepts were uploaded to OMOP, published at OHDSI's public and official vocabulary website, ATHENA [13], as shown in Figure 3. We removed 26 EDI concepts from among medical services because their codes (EDI IDs) were duplicated in the EDI device domain. They were already deprecated in the EDI vocabulary. The OHDSI vocabulary team assigned a unique OMOP identifier to all EDI concepts in February 2020.

We translated 273,449 Korean definitions of EDI concepts using the Google Cloud Translation API. After manual review, only 890 terms (0.33%) did not need further modification. The other 272,559 terms were retranslated with reference to the glossary. We present the results of the initial translation (without glossary) and second translation (referring glossary) in Figure 4. As seen in Figure 4, the translation procedures, including glossary constraints, achieved better performance for the meaning of abbreviations, medical terms, and descriptions.

The incorporation of the EDI vocabulary into the OMOP Standardized Vocabularies brought about three obvious improvements: (1) uniqueness and exclusivity of concepts, (2) hierarchies and relationships between concepts, (3) and a management system for vocabulary. The 11 criteria used to audit the current EDI vocabulary and the converted EDI were used more specifically. The criteria of concept orientation, coverage, non-semantic concept identifiers, and synonyms uniquely identified and mapped to relevant concepts were used to evaluate how unique and exclusive the concepts were. The systematic nature of the hierarchical structure was evaluated in terms of relation, multiple hierarchy, and formal definitions. The incorporated EDI vocabulary featured a more structured management system, evaluated in terms of concept permanence, version control, and multilanguage. For all criteria except compositionality, converted EDI showed a better quality index than the original EDI, as



Ablation of Liver Tumors-Laparoscopic-Microwave

Ablation of Liver Tum... ×
EDI ×

- ● DOMAIN ▼
- ● CONCEPT ▼
- ● CLASS ▼
- ● VOCAB ▲
  - □ DRG (159)
  - □ Death Type (3)
  - □ Domain (2)
  - ☑ EDI (101764)
  - □ EphMRA ATC (24)
  - □ Episode (1)
  - □ GCN_SEQNO (151)
  - □ GGR (2)
  - □ HCPCS (1793)
- ● VALIDITY ▼

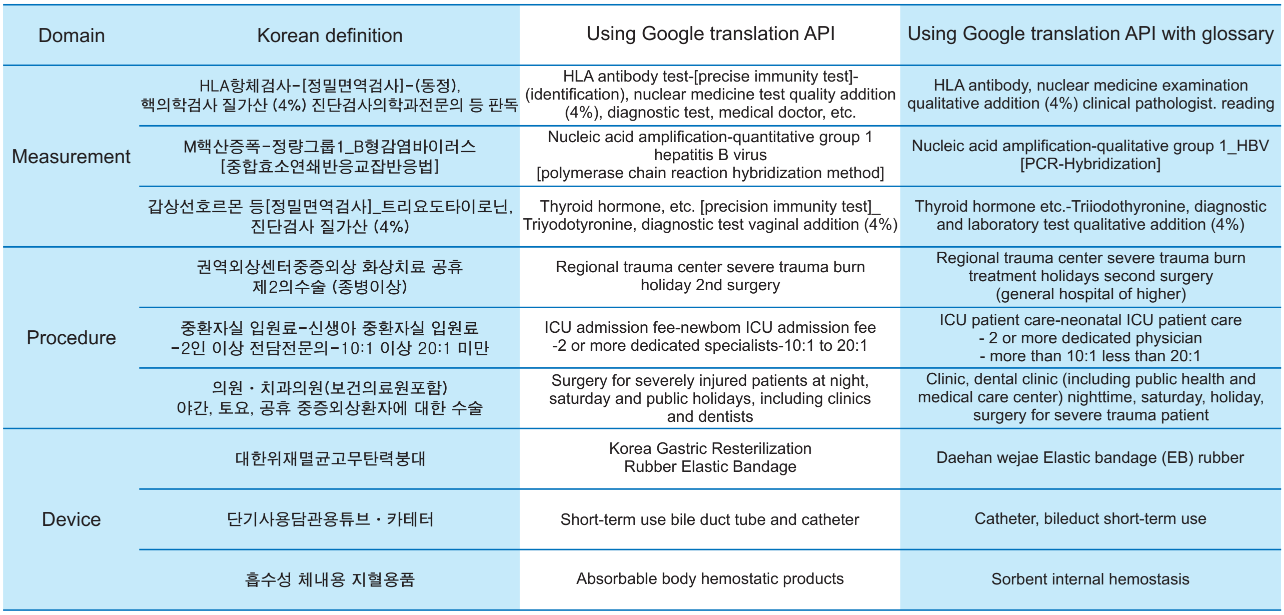
Show by 15 ▼ items Total 101,764 items 1 2 3 4 5 ··· 6785 >

ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
42361127	Q7285	Ablation of Liver Tumors-Laparoscopic-Microwave	Proc Hierarchy	Non-standard	Valid	Procedure	EDI
42165689	Q7285050	Ablation of Liver Tumors-Laparoscopic-Microwave,holiday	Procedure	Non-standard	Valid	Procedure	EDI
42272220	Q7285600	Ablation of Liver Tumors-Laparoscopic-Microwave,Neonate	Procedure	Non-standard	Valid	Procedure	EDI
42294163	Q7285010	Ablation of Liver Tumors-Laparoscopic-Microwave,Nighttime	Procedure	Non-standard	Valid	Procedure	EDI
42330927	Q7285001	Ablation of Liver Tumors-Laparoscopic-Microwave,second surgery	Procedure	Non-standard	Valid	Procedure	EDI
42275014	Q7285650	Ablation of Liver Tumors-Laparoscopic-Microwave,Neonate, holiday	Procedure	Non-standard	Valid	Procedure	EDI
42279438	Q7285610	Ablation of Liver Tumors-Laparoscopic-Microwave,Neonate, Nighttime	Procedure	Non-standard	Valid	Procedure	EDI
42215412	Q7285A00	Ablation of Liver Tumors-Laparoscopic-Microwave,Under 12 months	Procedure	Non-standard	Valid	Procedure	EDI
42189096	Q7285051	Ablation of Liver Tumors-Laparoscopic-Microwave,Holiday second surgery	Procedure	Non-standard	Valid	Procedure	EDI
42291597	Q7285601	Ablation of Liver Tumors-Laparoscopic-Microwave,Neonate second surgery	Procedure	Non-standard	Valid	Procedure	EDI
42298123	Q7285011	Ablation of Liver Tumors-Laparoscopic-Microwave,Nighttime second surgery	Procedure	Non-standard	Valid	Procedure	EDI

Ablation of Liver Tumors-Laparoscopic-Microwave,holiday

DETAILS	
Domain ID	Procedure
Concept Class ID	Procedure
Vocabulary ID	EDI
Concept ID	42165689
Concept code	Q7285050
Validity	Valid
Concept	Non-standard
Synonyms	간암에 실시하는 열치료술 [유도료 별도 산정]-복강경하-극초단파, 공휴일 Ablation of Liver Tumors-Laparoscopic-Microwave,holiday
Valid start	01-Jan-2019
Valid end	31-Dec-2099

Figure 3. Redacted EDI concepts were uploaded to OMOP, published at OHDSI's public and official vocabulary website, ATHENA (https://athena.ohdsi.org/). EDI: Electronic Data Interchange, OMOP: Observational Medical Outcomes Partnership, OHDSI: Observational Health Data Sciences and Informatics.



Domain	Korean definition	Using Google translation API	Using Google translation API with glossary
Measurement	HLA항체검사-[정밀면역검사]-(동정), 핵의학검사 질가산 (4%) 진단검사의학과전문의 등 판독	HLA antibody test-[precise immunity test]-(identification), nuclear medicine test quality addition (4%), diagnostic test, medical doctor, etc.	HLA antibody, nuclear medicine examination qualitative addition (4%) clinical pathologist. reading
	M핵산증폭-정량그룹1_B형감염바이러스 [중합효소연쇄반응교잡반응법]	Nucleic acid amplification-quantitative group 1 hepatitis B virus [polymerase chain reaction hybridization method]	Nucleic acid amplification-qualitative group 1_HBV [PCR-Hybridization]
	갑상선호르몬 등[정밀면역검사]_트리요도타이로닌, 진단검사 질가산 (4%)	Thyroid hormone, etc. [precision immunity test]_ Triyodotyronine, diagnostic test vaginal addition (4%)	Thyroid hormone etc.-Triiodothyronine, diagnostic and laboratory test qualitative addition (4%)
Procedure	권역외상센터중증외상 화상치료 공휴 제2의수술 (종병이상)	Regional trauma center severe trauma burn holiday 2nd surgery	Regional trauma center severe trauma burn treatment holidays second surgery (general hospital of higher)
	중환자실 입원료-신생아 중환자실 입원료 -2인 이상 전담전문의-10:1 이상 20:1 미만	ICU admission fee-newbom ICU admission fee -2 or more dedicated specialists-10:1 to 20:1	ICU patient care-neonatal ICU patient care - 2 or more dedicated physician - more than 10:1 less than 20:1
	의원・치과의원(보건의료원포함) 야간, 토요, 공휴 중증외상환자에 대한 수술	Surgery for severely injured patients at night, saturday and public holidays, including clinics and dentists	Clinic, dental clinic (including public health and medical care center) nighttime, saturday, holiday, surgery for severe trauma patient
Device	대한위재멸균고무탄력붕대	Korea Gastric Resterilization Rubber Elastic Bandage	Daehan wejae Elastic bandage (EB) rubber
	단기사용담관용튜브・카테터	Short-term use bile duct tube and catheter	Catheter, bileduct short-term use
	흡수성 체내용 지혈용품	Absorbable body hemostatic products	Sorbent internal hemostasis

Figure 4. The results of the initial translation (without glossary) and second translation (referring glossary). The translation procedures, including glossary constraints, achieved better performance for the meaning of abbreviations, medical terms, and descriptions.

shown in Table 2.

As previously stated, the criteria of concept orientation, non-semantic concept identifiers, coverage, and synonyms uniquely identified and mapped to relevant concepts were used to evaluate how unique and exclusive the concepts were. Concept orientation stipulates that a concept must correspond to a single meaning. Concept orientation is impaired in the current EDI vocabulary because it uses the same concept definitions for several concepts, despite the fact that they have different concept identifiers. In this case, concepts can be distinguished by a modifier for reimbursing the additional price of service. After incorporation into the OMOP vocabulary, converted EDI concepts gain unique concept definitions. The current EDI vocabulary uses semantic identifiers that have the advantage of having a meaning for each digit of the codes, which enables the easy identification of a single hierarchy (e.g., A3133 is parent code of A3133100, A3133200, and so on). However, if the rule for assigning concept codes changes for some reason, this convenience can become a constraint. Also, for vocabularies with multiple hierarchies, semantic identifiers can cause confusion [8]. Through incorporation into the OMOP vocabulary, a non-semantic identifier was assigned to every concept in the EDI vocabulary to meet the non-semantic

Table 2. Terminology evaluation criteria and explanation

Criteria		Explanation	EDI vocabulary	EDI in OMOP vocabulary
Uniqueness and exclusivity of the concept	Concept orientation	A concept must be linked with only one term	△	○
	Non-semantic concept identifiers	There must be a unique code representing a concept	×	○
	Coverage	The domain covered by the terminology system must be consistent and obvious	○	○
	Synonyms uniquely identified and mapped to relevant concepts	Synonyms, including abbreviations, are managed by unique identifiers, and related concepts are mapped	×	○
Hierarchies and relationships between concepts	Relation	The relation of each concept should be defined	×	○
	Multiple hierarchy	A concept can have multiple hierarchies	×	△
	Formal definition	Having a structure and definition that can be indexed and processed by computer	×	○
	Compositionality	Terms can be separated into atomic units and have compositional extensibility	×	×
Management system for vocabulary	Concept permanence	Even if the used term is updated, the previously used term should not be deleted	×	○
	Version control	When terminology is updated, version information, including changes, must be specified	×	○
	Multi-language	The terminology system supports multiple languages	△	○

EDI: Electronic Data Interchange, OMOP: Observational Medical Outcomes Partnership.

concept identifier criterion. We classified the medical service domain of the current EDI vocabulary as measurements and procedures. Although the converted EDI vocabulary has more specific domains, both have consistent and obvious coverage. Regarding the “synonyms uniquely identified and mapped to relevant concepts” criterion, the current EDI vocabulary does not have the structure of concept relationship, and it contains some duplicated identifiers. However, concepts in the converted EDI vocabulary have defined relationships between associated concepts and unique Korean definitions as concept synonyms of English definitions that meet this criterion.

The systematic nature of the hierarchical structure was evaluated in terms of relation, multiple hierarchies, formal definition, and compositionality criteria. Relation refers to the existing connections between related concepts. The current EDI does not maintain any relation between concepts, whereas the converted EDI has the structure of concept relation, allowing related concepts in other vocabularies to be identified. The hierarchy of concepts is established by defining the horizontal/vertical relationships of concepts. We constructed a single vertical hierarchy in the converted EDI vocabulary. However, it does not fully meet the multiple hierarchy criteria. The formal definition refers to a structure with concept relations that can be indexed and processed by a computer. The current EDI vocabulary lacks formal definition because even Korean definitions of EDI concepts vary across the versions of the EDI that have been released. Furthermore, the EDI vocabulary does not provide a system to search for related concepts based on the definitions of concepts. The converted EDI is available at the official OMOP vocabulary website, ATHENA [13], where users can easily search for related concepts using formal English definitions. Compositionality refers to the fact that composite concepts can be divided into simple atomic concepts. This provides an intuitive understanding of complex concepts, but the current and converted EDI vocabularies do not meet this criterion.

The incorporated EDI vocabulary featured a more structured management system when it is evaluated according to concept permanence, version control, and multi-language criteria. Concept permanence means that expired or modified concepts and identifiers remain permanently. The current EDI vocabulary removes the expired concepts and reassigns the deprecated identifiers to newly added concepts. The converted EDI vocabulary maintains the expired concepts and identifiers. Version control is the corollary of concept permanence. The converted EDI vocabulary enables versioning through storing metadata of the start and the expiry date for each concept. In addition, the current EDI provides English definitions only for some concepts, whereas the converted EDI provides unique English definitions for all concepts.

IV. Discussion

We audited the Korean EDI as a controlled medical vocabulary in use in the Korean EHR system. By incorporating the EDI vocabulary into the OMOP Standardized Vocabulary, we enhanced many aspects of a controlled vocabulary, such as concept permanence, consistency, versioning, hierarchy, relations between concepts, formal definitions, unique and non-semantic identifiers, as well as expressive Korean and English definitions of concepts, while maintaining the EDI's coverage. As a controlled vocabulary, the EDI in the OMOP vocabulary can provide a cohort database with unified terms and normalized concepts to researchers with similar research purposes. We also developed and deployed an open-source R package to automate this procedure.

The objective of this study was not to investigate errors in the EDI vocabulary. Rather, the ultimate aim of this study was to further improve the EDI vocabulary for a controlled and standardized vocabulary system. The EDI vocabulary itself was created for the purely administrative purpose of facilitating nationwide insurance. It was not designed as a comprehensive medical ontology. Nonetheless, the EDI vocabulary has become the *de facto* vocabulary for observational medical research in Korea because of the rapid expansion of the secondary use of Korean EHR and the administrative claims database for real-world evidence [14,15]. Unlike the EDI, the converted EDI concepts in the OMOP Standardized Vocabularies were assigned unique identifiers, and they may have exclusive definitions. In the OMOP, each concept corresponds to no more than one meaning and is exclusive, resulting in better concept orientation.

This study provides significant advantages for big data analysis when using a Korean medical database. First, it helps to build a standard process for transforming Korean observational databases into OMOP-CDM. We recommend storing the OMOP concept IDs of EDI concepts in the "_SOURCE_CONCEPT_ID" fields of drug exposure, procedure occurrence, device exposure, and measurement tables. Then, EDI concept-based collaborative research can be performed across Korean databases without the need for further vocabulary mapping. Second, it may enhance the transparency and reproducibility of Korean medical research. Until now, most studies using the Korean Administrative Claims Database have not provided actual EDI identifiers because no English documentation for EDI concepts has existed [16]. Because our study provides formal English definitions and a hierarchy of EDI concepts, it may precipitate the reporting of EDI identifiers in scientific papers, enhancing the reproducibility of research. Third, our study paves the way for international collaborative research using Korean databases. In response to the coronavirus disease 2019 (COVID-19) pandemic, HIRA launched a global research collaboration project with clinical data from Korean patients with COVID-19 on March 27, 2020 [17]. Although there was no official English document describing the dataset's medical vocabulary, all of the necessary information for KCD-7 and the EDI vocabulary was accessible through the ATHENA web portal [13].

As originally intended, incorporation of the EDI vocabulary into the OMOP Standardized Vocabularies provides the infrastructure for standard mapping. As of September 2020, we had published corresponding standard concepts for 37,869 EDI procedure concepts and 675 measurement concepts [18]. Most of the OMOP standard concepts for procedures and measurements were derived from SNOMED-CT and the LOINC (Logical Observation Identifiers Names and Codes) vocabulary, respectively.

This study had several limitations. First, only vocabulary current as of October 2019 has been incorporated into OMOP. To provide updates, Korean definitions of new terms should be translated into English by a human translator. By August 2020, a total of 447 EDI codes for Korean synonyms had been changed, 10,320 codes had been newly added, and 7,873 codes had been deprecated. Regular updates of changes in the EDI vocabulary to the OMOP vocabulary should be conducted going forward. Second, the quality of English definitions of EDI concepts has not been fully evaluated by professional medical staff. All information is publicly available [13], and the overall quality can be improved through open discussion [19]. Third, we used the list of EDI concept list released on the HIRA website, but it does not include concepts that had expired before October 2019.

By incorporating the EDI vocabulary into the OMOP Standardized Vocabularies, Korean medical terms can become standardized. This research developed a promising approach to mapping Korean medical information into a global standard system of terminology, but comprehensive official vocabulary mapping remains to be done in the future.

https://doi.org/10.4258/hir.2021.27.1.29

Conflict of Interest

Dr. Dymshyts is employed by Odysseus Data Services, Inc. Dr. Reich is an employee of IQVIA. There are no financial incentives related to the work described here. The other authors declare no competing financial interests. None of the sponsors or funders above, including Odysseus and IQVIA, had input in the design, execution, interpretation of results, or decision to publish.

Rae-Woong Park is an editor of Healthcare Informatics Research; however, he did not involve in the peer reviewer selection, evaluation, and decision process of this article. The rest of the authors declare no conflict of interest.

Acknowledgments

This work was supported by the Health Insurance Review & Assessment Service (HIRA). The views expressed are those of the authors and not necessarily those of the HIRA.

ORCID

Yeonchan Seong (http://orcid.org/0000-0003-4201-7161)
Seng Chan You (http://orcid.org/0000-0002-5052-6399)
Anna Ostropolets (http://orcid.org/0000-0002-0847-6682)
Yeunsook Rho (http://orcid.org/0000-0001-5339-1082)
Jimyung Park (http://orcid.org/0000-0002-6998-2546)
Jaehyeong Cho (http://orcid.org/0000-0001-7213-5033)
Dmitry Dymshyts (http://orcid.org/0000-0001-8718-0013)
Christian G. Reich (http://orcid.org/0000-0002-3641-055X)
Yunjung Heo (http://orcid.org/0000-0001-5708-1428)
Rae Woong Park (http://orcid.org/0000-0003-4989-3287)

References

1. Park HA, Kim HY, Min YH. Use of clinical terminology for semantic interoperability of electronic health records. J Korean Am Med Assoc 2012;55(8):720-8.
2. Healthcare Information Standard [Internet]. Seoul, Korea: Korea Health Information Service; 2017 [cited at 2021 Jan 26]. Available from: https://www.hins.or.kr/cmm/main/mainPage.do.
3. Kim M. A study on a comparison of diagnostic domain between SNOMED CT and Korea standard terminology of medicine. Int J Database Theory Appl 2015;9(11):49-60.
4. Jung BK, Kim J, Cho CH, Kim JY, Nam MH, Shin BK, et al. Report on the Project for Establishment of the Standardized Korean Laboratory Terminology Database, 2015. J Korean Med Sci 2017;32(4):695-9.
5. Hwang EJ, Park HA, Sohn SK, Lee HB, Choi HK, Ha S, et al. Mapping Korean EDI medical procedure code to SNOMED CT. Stud Health Technol Inform 2019;264:178-82.
6. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-8.
7. 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;64:333-41.
8. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.
9. EdiToOmop: the package to convert Korean EDI code to the OMOP vocabulary [Internet]. San Francisco (CA): GitHub Inc.; 2020 [cited at 2021 Jan 26]. Available from: https://github.com/OHDSI/EdiToOmop.
10. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. J Am Med Inform Assoc 1998;5(6):503-10.
11. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13(3):277-88.
12. Lee KC. A study on enhancing terminology & its major issues [dissertation]. Seoul, Korea: Yonsei University; 2012.
13. ATHENA: OHDSI vocabularies repository [Internet]. Bethesda (MD): Observational Health Data Sciences and Informatics; 2018 [cited at 2021 Jan 26]. Available from: http://athena.ohdsi.org.
14. Han SH, Lee MH, Kim SG, Jeong JY, Lee BN, Choi MS, et al. Implementation of medical information exchange system based on EHR standard. Health Inform Res 2010;16(4):281-9.
15. Kim TG, Kwon O, Shin YS, Sung JH, Koh JS, Kim BT. Endovascular treatments performed collaboratively by the Society of Korean Endovascular Neurosurgeons

members: a nationwide multicenter survey. J Korean Neurosurg Soc 2019;62(5):502-18.

16. You SC, An MH, Yoon D, Ban GY, Yang PS, Yu HT, et al. Rate control and clinical outcomes in patients with atrial fibrillation and obstructive lung disease. Heart Rhythm 2018;15(12):1825-32.

17. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahao MT, et al. Deep phenotyping of 34,128 patients hospitalised with COVID-19 and a comparison with 81,596 influenza patients in America, Europe and Asia: an international network study. medRxiv 2020 [Epub]. https://doi.org/10.1101/2020.04.22.20074336.

18. Korean OMOP Vocabulary [Internet]. San Francisco (CA): GitHub Inc.; 2019 [cited at 2021 Jan 26]. Available from: https://github.com/ohdsi-korea/OmopVocabularyKorea.

19. OHDSI-Korea [Internet]. San Francisco (CA): GitHub Inc.; 2019 [cited at 2021 Jan 26]. Available from: https://github.com/ohdsi-korea.