**Original Article**

Healthc Inform Res. 2022 July;28(3):240-246.
https://doi.org/10.4258/hir.2022.28.3.240
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

# Effectiveness of the Use of Standardized Vocabularies on Epilepsy Patient Cohort Generation

**Hyesil Jung[1], Ho-Young Lee[2], Sooyoung Yoo[1], Hee Hwang[3], Hyunyoung Baek[1]**

[1]Healthcare ICT Research Center, Seoul National University Bundang Hospital, Seongnam, Korea
[2]Department of Nuclear Medicine, Seoul National University Bundang Hospital, Seongnam, Korea
[3]Kakao Healthcare Company-In-Company, Seongnam, Korea

**Objectives:** This study investigated the effectiveness of using standardized vocabularies to generate epilepsy patient cohorts with local medical codes, SNOMED Clinical Terms (SNOMED CT), and International Classification of Diseases tenth revision (ICD-10)/Korean Classification of Diseases-7 (KCD-7). **Methods:** We compared the granularity between SNOMED CT and ICD-10 for epilepsy by counting the number of SNOMED CT concepts mapped to one ICD-10 code. Next, we created epilepsy patient cohorts by selecting all patients who had at least one code included in the concept sets defined using each vocabulary. We set patient cohorts generated by local codes as the reference to evaluate the patient cohorts generated using SNOMED CT and ICD-10/KCD-7. We compared the number of patients, the prevalence of epilepsy, and the age distribution between patient cohorts by year. **Results:** In terms of the cohort size, the match rate with the reference cohort was approximately 99.2% for SNOMED CT and 94.0% for ICD-10/KDC7. From 2010 to 2019, the mean prevalence of epilepsy defined using the local codes, SNOMED CT, and ICD-10/KCD-7 was 0.889%, 0.891% and 0.923%, respectively. The age distribution of epilepsy patients showed no significant difference between the cohorts defined using local codes or SNOMED CT, but the ICD-9/KCD-7-generated cohort showed a substantial gap in the age distribution of patients with epilepsy compared to the cohort generated using the local codes. **Conclusions:** The number and age distribution of patients were substantially different from the reference when we used ICD-10/KCD-7 codes, but not when we used SNOMED CT concepts. Therefore, SNOMED CT is more suitable for representing clinical ideas and conducting clinical studies than ICD-10/KCD-7.

**Keywords:** Systematized Nomenclature of Medicine, Terminology, Cohort Studies, Epilepsy, International Classification of Diseases

**Submitted:** September 28, 2021
**Revised:** March 23, 2022
**Accepted:** April 24, 2022

**Corresponding Author**
Ho-Young Lee
Department of Nuclear Medicine, Seoul National University Bundang Hospital, 82 Gumi-ro 173 Beon-gil, Bundang-gu, Seongnam 13620, Korea. Tel: +82-31-787-2938, E-mail: debobkr@snubh.org (https://orcid.org/0000-0001-6518-0602)



## I. Introduction

The goal of adopting healthcare information technology is to optimize the collection, sharing, and utilization of data generated in the field of healthcare. Utilizing clinical data efficiently requires interoperability. Specifically, semantic interoperability, which preserves the semantics of the exchanged data, can be achieved using standardized terminologies. The standardization of clinical data makes it possible to efficiently conduct health information exchange and multinational network studies.

SNOMED Clinical Terms (SNOMED CT) is the most

widely used system of clinical terminology with fine granularity and an extensive hierarchy; it is also increasingly used for clinical data entry and retrieval. SNOMED CT is used in 40 member countries and by more than 5,000 individuals and organizations around the world [1]. In August 2020, Korea became the 39th member of SNOMED International, and domestic healthcare institutions are actively working to achieve semantic interoperability and standardization of health data using SNOMED CT. SNOMED CT has been used in Korean national IT initiatives, such as the program for the certification of Electronic Medical Records (EMR) systems, for health information exchange, for data-driven hospitals, and for national registries (e.g., the Cancer Registration and Statistics Program) [2].

Seoul National University Bundang Hospital (SNUBH) has been using standard terminologies, including SNOMED CT, for semantic interoperability and data utilization since it was founded in 2003. Specifically, all diagnoses, chief complaints, and surgical procedure codes were standardized using SNOMED CT and International Classification of Diseases ninth revision with clinical modification (ICD-9-CM). In addition, significant numbers of clinical observation records (e.g., vital signs and pain scores), radiology and pathology reports, and laboratory test results were mapped to the corresponding Logical Observation Identifier Names and Codes.

However, there are concerns about information loss whenever a mapping is performed [3]. Mapping concepts from one taxonomy to another can create semantic inconsistency due to hierarchy incongruence [4,5]. There is a difference in granularity between terminologies; for example, a source concept may be either too specific or too general to be directly mapped to a target concept [6]. Despite these concerns, Reich et al. [4] showed that, although there are vocabulary differences in mapping from ICD-9-CM to SNOMED CT and these differences cause differences in cohorts, studies that used these mappings showed minimal differences compared with those of the original studies. Hripcsak et al. [3] also showed that mapping data from source ICD billing codes to SNOMED CT codes had only a very small effect on the generated patient cohorts. However, another study showed substantial inconsistencies and disagreements between patient cohorts generated by standardized vocabularies and original codes across network sites [7].

In this study, we evaluated the effectiveness of the use of standardized vocabularies to generate epilepsy patient cohorts with local medical codes, SNOMED CT, and International Classification of Diseases tenth revision (ICD-10)/Korean Classification of Diseases-7 (KCD-7) and compared the cohorts in terms of the number and age distribution of the patients by year.

## II. Methods

To compare the granularity between SNOMED CT and ICD-10 for epilepsy, we counted the number of SNOMED CT concepts mapped to one ICD-10 code in the SNOMED CT to an ICD-10 map released by SNOMED International and the World Health Organization [8]. Next, we created and compared patient cohorts by selecting all patients who had at least one code that was included in the concept sets of the local codes, SNOMED CT, and ICD-10/KCD-7. We used patients' primary and secondary diagnosis codes from inpatient visits, outpatient visits, and emergency room visits.

### 1. Concept Set of Local Codes

To establish a concept set for epilepsy, we used diagnosis codes defined in a previous study [9]. Moreover, we added 26 local codes with "%epilep%" in the code name to include patients with "status epilepticus," "posttraumatic epilepsy," and "acquired epileptic aphasia." We determined whether the added local codes were related to epilepsy by examining the supertype ancestor of the SNOMED CT concepts mapped to the local codes.

### 2. Concept Sets of SNOMED CT and ICD-10/KCD-7

We created a concept set of SNOMED CT including the 84757009 |Epilepsy (disorder)| concept and its descendent concepts in ATLAS, the Observational Health Data Sciences and Informatics (OHDSI) open-source software that facilitates the design and execution of analyses of standardized, observational data in the common data model (CDM) format. SNUBH ATLAS uses SNOMED CT International (released on April 1, 2020) from the Observational Medical Outcomes Partnership (OMOP) Standardized vocabularies. According to a report by the International League Against Epilepsy Task Force on ICD codes in epilepsy [10], we created a concept set of ICD-10, including G40.X (epilepsy) G41.X (status epilepticus), and F80.3 (Landau-Kleffner syndrome). Korea uses codes from the KCD-7, which is a Korean version of the ICD-10 that represents detailed epilepsy-related information, such as patients with or without intractable epilepsy, using two decimal places (e.g., G40.30, generalized idiopathic epilepsy and epileptic syndromes without intractable epilepsy). Therefore, we included the G40 codes with two decimal places in the concept set.

### 3. Patient Cohorts

We applied the concept sets to the OMOP CDM database of SNUBH, which contains data from 2 million patients obtained between May 2003 and July 2019. We defined cohorts of patients as those who had at least one code from a concept set in their records. We set patient cohorts generated by local codes as the reference to evaluate patient cohorts by SNOMED CT and ICD-10/KCD-7. We counted the number of patients included in or missing from the patient cohorts generated by SNOMED CT and ICD-10/KCD-7 compared to the reference. We also compared the prevalence of epilepsy and age distribution between patient cohorts by year. Patients with epilepsy were extracted and classified by year, according to condition_start_date in a condition_occurrence table. Figure 1 shows an overview of the study with the created and compared patient cohorts based on local codes, SNOMED CT, and ICD-10/KCD-7.

This study received approval from the SNUBH Institutional Review Board (No. X-2109-711-904) and was performed in accordance with the relevant guidelines and regulations of the IRB.

## III. Results

### 1. Granularity between SNOMED CT and ICD-10/KCD-7

Figure 2 shows epilepsy-related codes in the local codes, SNOMED CT, and ICD-10/KCD-7. Of 174 local codes for epilepsy, 133 codes were used to map diagnoses of epilepsy in EMRs. Although we used 298 SNOMED CT concepts to define the patient cohort with epilepsy, we found that only 52 concepts were used to map epilepsy-related diagnoses in EMRs. Of the 36 ICD-10/KCD-7 codes, only 21 codes (G40.0, G40.00, G40.01, G40.1, G40.2, G40.20, G40.21, G40.3, G40.30, G40.31, G40.4, G40.5, G40.6, G40.7, G40.8, G49.9, G41.1, G41.2, G41.8, G41.9, and F80.3) were mapped and used in EMRs.

One ICD-10 code was mapped to multiple SNOMED CT concepts (up to 84 concepts) in the SNOMED CT to ICD-10 map, as shown in Table 1. ICD-10 has complex concepts (i.e., G40.1, localization-related symptomatic epilepsy and epileptic syndromes with simple partial seizures) that exist only as separate codes (e.g., 117891000119100 |Simple partial seizure (disorder)|, 230381009 |Localization-related epilepsy (disorder)|) in SNOMED CT.

### 2. Size of the Patient Cohorts

Table 2 shows the size differences between the patient cohorts generated by local codes, SNOMED CT, and ICD-10/KCD-7. For the reference group, the epilepsy-related local codes led to the extraction of 11,141 patients from 2003 to 2019, of whom 53.2% were male. The size of the patient cohort generated using the SNOMED CT concepts was 1.007-fold larger than that of the reference. Of the 11,220 patients, 99.2% were matched with the reference.

In addition, the patient cohort generated by ICD-10/KCD-7 included 6% more patients than the reference cohort, and the match rate with the reference was approximately 94.0%. Almost 0.4% of the patients included in the reference cohort were missing from the cohort generated by ICD-10/KCD-7.

### 3. Prevalence of Epilepsy and Age Distribution by Year

Figure 3 shows the prevalence of epilepsy by year according to vocabularies used to generate the patient cohorts. From 2010 to 2019, the mean prevalence of epilepsy was 0.889% when we used local codes to define epilepsy. In the SNOMED CT and ICD-10/KCD-7 cohorts, the mean prevalence of epilepsy was 0.891% and 0.923%, respectively.

Figure 4 shows the age distribution of epilepsy patients whose data were extracted by using local codes (reference)

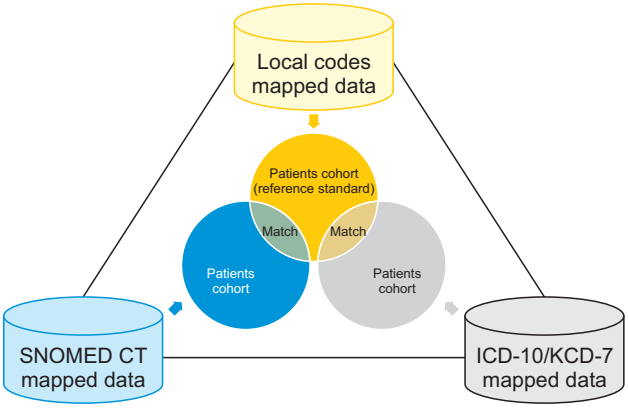


Figure 1. Design of the study. SNOMED CT: SNOMED Clinical Terms, ICD-10: International Classification of Diseases tenth revision, KCD-7: Korean Classification of Diseases-7.

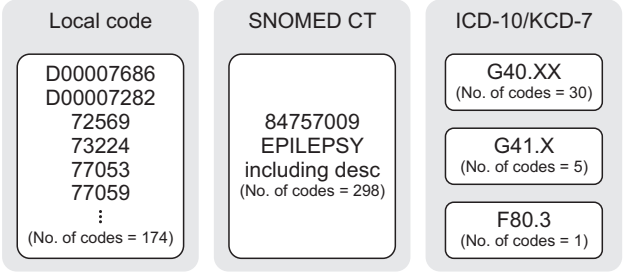


Figure 2. Local, SNOMED CT, and ICD-10/KCD-7 codes for epilepsy. SNOMED CT: SNOMED Clinical Terms, ICD-10: International Classification of Diseases tenth revision, KCD-7: Korean Classification of Diseases-7.


https://doi.org/10.4258/hir.2022.28.3.240

Table 1. Number of SNOMED CT concepts mapped to a single ICD-10 code for epilepsy

| ICD-10 code | ICD-10 name | Number of SCT codes mapped |
|---|---|---|
| G40.0 | Localization-related (focal) (partial) idiopathic epilepsy and epileptic syndromes with seizures of localized onset | 24 |
| G40.1 | Localization-related (focal) (partial) symptomatic epilepsy and epileptic syndromes with simple partial seizures | 59 |
| G40.2 | Localization-related (focal) (partial) symptomatic epilepsy and epileptic syndromes with complex partial seizures | 34 |
| G40.3 | Generalized idiopathic epilepsy and epileptic syndromes | 84 |
| G40.4 | Other generalized epilepsy and epileptic syndromes | 38 |
| G40.5 | Special epileptic syndromes | 15 |
| G40.6 | Grand mal seizures, unspecified (with or without petit mal) | 4 |
| G40.7 | Petit mal, unspecified, without grand mal seizures | 1 |
| G40.8 | Other epilepsy | 49 |
| G40.9 | Epilepsy, unspecified | 36 |
| G41.0 | Grand mal status epilepticus | 2 |
| G41.1 | Petit mal status epilepticus | 4 |
| G41.2 | Complex partial status epilepticus | 3 |
| G41.8 | Other status epilepticus | 5 |
| G41.9 | Status epilepticus, unspecified | 11 |
| F80.3 | Acquired aphasia with epilepsy [Landau-Kleffner] | 1 |

SNOMED CT: SNOMED Clinical Terms, ICD-10: International Classification of Diseases tenth revision, SCT: SNOMED CT.

Table 2. Number of patients in the cohorts generated by local codes, SNOMED CT, and ICD-10/KCD-7

| | Local code (RS) | SNOMED CT | ICD-10/KCD-7 |
|---|---|---|---|
| Total | 11,141 | 11,220 (100) | 11,812 (100) |
| Match with RS | - | 11,132 (99.2) | 11,099 (94.0) |
| False positive[a] | - | 88 (0.8) | 713 (6.0) |
| False negative[b] | - | 9 | 42 |

Values are presented as number of patients (%).
RS: reference standard, SNOMED CT: SNOMED Clinical Terms, ICD-10: International Classification of Diseases tenth revision, KCD-7: Korean Classification of Diseases-7.
[a]Number of patients included in the cohort generated by SNOMED CT or ICD-10/KCD-7 but missing from the reference standard.
[b]Number of patients included in the reference standard but missing from the cohort generated by using SNOMED CT or ICD-10/KCD-7.

from 2003 to 2019. Except in 2003, the proportion of patients under the age of 20 years was large in all age groups. Upon comparing the age distribution of epilepsy patients among the cohorts generated using the local codes, SNOMED CT, and ICD-10/KCD-7, the local codes and SNOMED CT showed no significant difference in the age distribution of epilepsy patients. However, the age distribution of patients with epilepsy showed a substantial gap between the data extracted by the local codes and ICD-10/KCD-7, as shown in Figure 5. Notably, the proportion of pediatric patients with epilepsy under the age of 10 years was greater in the patient cohort generated by ICD-10/KCD-7 than in the patient cohorts generated by local codes or SNOMED CT.

## IV. Discussion

To analyze the effects of data standardization (vocabulary mapping), previous studies compared patient cohorts [3,6,7] and evaluated the prevalence of specific health outcomes [4] and estimates of drug-heath outcome associations [4] across the mapped vocabularies in various databases.

This study evaluated the effect of data standardization in terms of generating a cohort of patients with epilepsy. We considered the patient cohort created using local codes as the reference and compared it with cohorts generated by

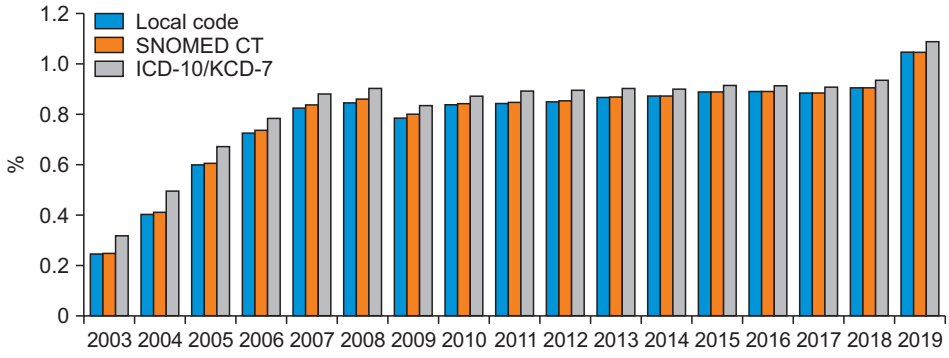


Figure 3. Prevalence of epilepsy by year in the three cohorts. SNOMED CT: SNOMED Clinical Terms, ICD-10: International Classification of Diseases tenth revision, KCD-7: Korean Classification of Diseases-7.

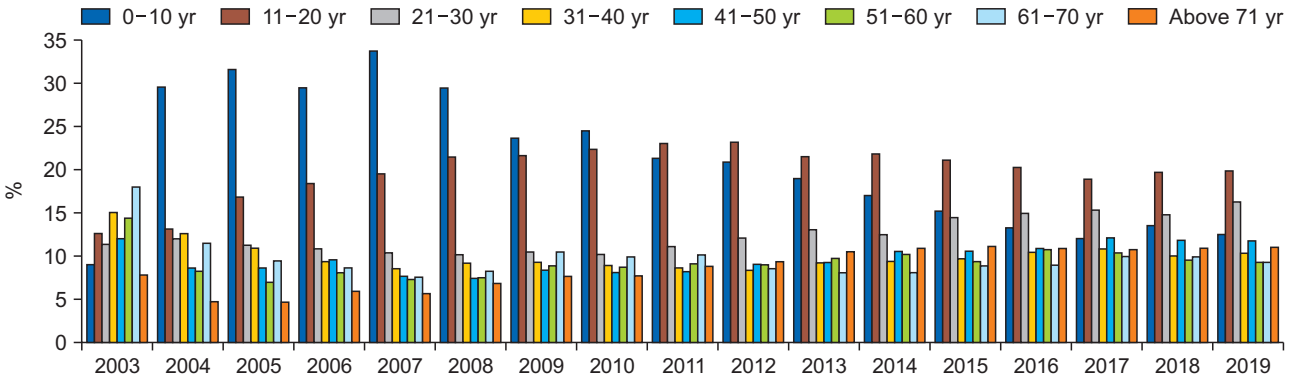


Figure 4. Age distribution of the epilepsy patient cohorts by year (reference standard).

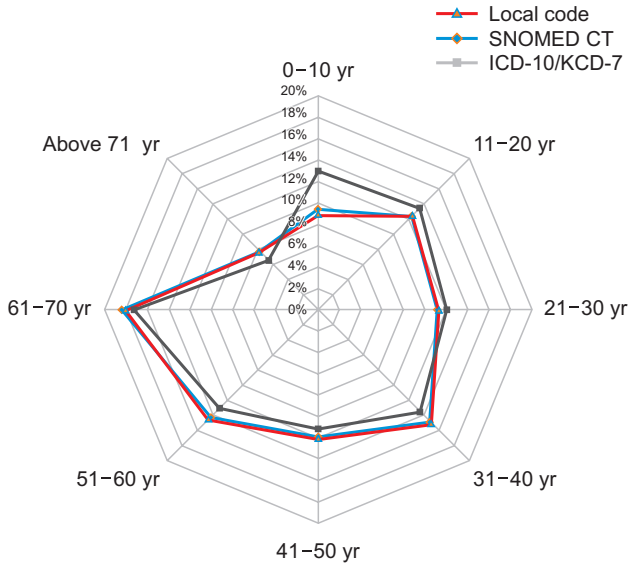


Figure 5. Difference in the age distribution of epilepsy patients by vocabularies in 2003. SNOMED CT: SNOMED Clinical Terms, ICD-10: International Classification of Diseases tenth revision, KCD-7: Korean Classification of Diseases-7.

SNOMED CT and ICD-10/KCD-7 in terms of the number and age distribution of the patients by year.

SNOMED CT is designed for direct use by healthcare providers during the process of care, whereas ICD-10 is designed for use by medical coders once an episode of care is completed. ICD-10 is a classification system that consists of groups of mutually exclusive categories for data aggregation. SNOMED CT, in contrast, is a health terminology that satisfies the requirements for reference terminologies, including concept orientation, formal definitions, poly-hierarchy, and multiple granularities [11]. Since SNOMED CT allows coding at any level of granularity that is appropriate for the clinical situation using a sub-type relationship, it is suited for documenting clinical information or ideas within EMRs. The SNOMED CT hierarchy allows facile incorporation of new concepts and increased granularity, eliminating the need to rely on ambiguous classifications such as NOS (not otherwise specified) and NEC (not elsewhere classifiable) codes, as are used in ICD-10 codes [12]. This increased granularity also benefits clinical research.

Of the patients included in the cohort generated by SNOMED CT, 88 patients were excluded from the reference. The local diagnoses of these patients had spelling errors, such as "benign myoclonic epilopsy [sic] in infancy," "benign myoclonic epilopsy [sic] in infancy, not intractable," and some had local names without "%epilep%" (e.g., benign neonatal familial convulsions, seizure with specific mode of precipitation); hence, we could not include these codes in the reference. In other words, SNOMED CT detected hidden patients with epilepsy that could not be identified using local


https://doi.org/10.4258/hir.2022.28.3.240

codes and included them in the cohort.

The epilepsy cohort generated by local codes contained patients diagnosed with situation-related seizures, but these patients were not included in the SNOMED CT cohort. Since 230431001 |Situation-related seizures (disorder)| concept is a sibling of 84757009 |Epilepsy (disorder)| concept in the SNOMED CT hierarchy, nine patients with situation-related seizures were not included in the cohort generated by 84757009 |Epilepsy (disorder)| concept and its descendants.

In addition, 42 patients included in the reference were missed from the cohort generated by the G40.XX, G41.X, and F80.3 codes, since the local codes for "hippocampal sclerosis" and "posttraumatic epilepsy" have been mapped to G37.9 (demyelinating disease of central nervous system, unspecified) and T90.5 (sequelae of intracranial injury), respectively. Most of the 713 patients included in the cohort generated by ICD-10/KCD-7 codes were patients with seizures, including epileptic seizures, simple/complex partial seizures, grand/petit mal seizures, and generalized tonic-clonic seizures. As the ICD-10/KCD-7 allows complex concepts to be encoded, patients with diagnoses and symptoms other than epilepsy might be included. Thus, the use of ICD-10/KCD-7 may hinder the homogeneity of study subjects when organizing the cohort.

Although we did not analyze the effects of data standardization on specific study outcomes (e.g., estimates of drug-disease associations), as was done in the study of Reich et al. [4], we found substantial differences in the number and age distribution of patients from the reference when we used ICD-10/KCD-7 codes, not SNOMED CT concepts, to generate a targeted patient cohort. This finding indicates that SNOMED CT is more suitable for representing clinical concepts or ideas than ICD-10/KCD-7 and is beneficial for clinical studies. Moreover, we evaluated the quality of mapping between vocabularies.

Our study has several limitations. First, we only generated patient cohorts with epilepsy at a single healthcare institution to evaluate the effect of data standardization. Second, we only used diagnosis codes to generate patient cohorts with epilepsy. Hripcsak et al. [3] used public phenotypes from the eMERGE initiative (https://phekb.org/phenotypes) to test the effect of mapping diagnosis codes from ICD-9-CM/ICD-10-CM to SNOMED CT on patient cohorts. The eMERGE initiative was chosen because the phenotype definitions were validated and the phenotypes were explained in each case, thereby allowing us to assess intent. Therefore, we could add the inclusion criterion of one or more prescriptions of antiepileptic drugs to identify subjects with epilepsy, as defined by the eMERGE initiative [13]. Third, we searched the diagnosis name with "%epilep%" to define epilepsy-related local codes. Thus, patients with the codes for "benign neonatal familial convulsions" and "benign myoclonic epilopsy [sic] in infancy" were missing from the reference. Fourth, the conversion from one vocabulary to another depends on the quality of the mapping tables and the mapping skills of the medical coders [4]. Thus, the mapping results can vary according to the mapping purpose and institution. In-depth understanding and training for standard terminologies are required to improve the quality of mapping between vocabularies. Fifth, the concept set used to define a phenotype depends on the version of a vocabulary. There are a total of 249 SNOMED CT concepts for epilepsy and its descendants in SNOMED CT International released on February 28, 2022. Since we used the 298 concepts for epilepsy in SNOMED CT International released on April 1, 2020, there may be differences in the cohort composition according to the SNOMED CT version.

We plan to expand our empirical research by phenotyping other health outcomes of interest (e.g., heart failure, diabetes mellitus) or identifying the effect of data standardization on estimates of drug-health outcomes associations, as in previous studies [3,4].

## Conflict of Interest

No potential conflict of interest relevant to this article was reported.

## Acknowledgments

This work was supported by a grant of the Medical Data-Driven Hospital support project through the Korea Health Information Service (KHIS), funded by the Ministry of Health & Welfare, Republic of Korea. Also, This work was supported by the National Research Foundation of Korea funded by the Ministry of Science and Information and Communication Technologies (grant NRF-2021R1A2C1091261).

## ORCID

Hyesil Jung (https://orcid.org/0000-0002-8346-9343)
Ho-Young Lee (https://orcid.org/0000-0001-6518-0602)
Sooyoung Yoo (https://orcid.org/0000-0001-8620-4925)
Hee Hwang (https://orcid.org/0000-0002-7964-1630)
Hyunyoung Baek (https://orcid.org/0000-0003-0810-9396)

## References

1. SNOMED International. Members [Internet]. London, UK: SNOMED International; c2022 [cited at 2022 Jun 30]. Available from: https://www.snomed.org/our-stakeholders/members.
2. Park HA, Yu SJ, Jung H. Strategies for adopting and implementing SNOMED CT in Korea. Healthc Inform Res 2021;27(1):3-10. https://doi.org/10.4258/hir.2021.27.1.3
3. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. J Am Med Inform Assoc 2018;25(12):1618-25. https://doi.org/10.1093/jamia/ocy124
4. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012;45(4):689-96. https://doi.org/10.1016/j.jbi.2012.05.002
5. Observational Health Data Sciences and Informatics. The book of OHDSI [Internet]. [place unknown]: Observational Health Data Sciences and Informatics; 2021 [cited at 2022 Jun 30]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/.
6. Papez V, Moinat M, Payralbe S, Asselbergs FW, Lumbers RT, Hemingway H, 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
7. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform 2019;96:103253. https://doi.org/10.1016/j.jbi.2019.103253
8. SNOMED International. SNOMED CT to ICD-10 mapping technical guide [Internet]. London, UK: SNOMED International; c2020 [cited at 2022 Jun 30]. Available from: http://snomed.org/icd10map.
9. Kim H, Yoo S, Jeon Y, Yi S, Kim S, Choi SA, et al. Characterization of anti-seizure medication treatment pathways in pediatric epilepsy using the electronic health record-based common data model. Front Neurol 2020;11:409. https://doi.org/10.3389/fneur.2020.00409
10. Jette N, Beghi E, Hesdorffer D, Moshe SL, Zuberi SM, Medina MT, et al. ICD coding for epilepsy: past, present, and future: a report by the International League Against Epilepsy Task Force on ICD codes in epilepsy. Epilepsia 2015;56(3):348-55. https://doi.org/10.1111/epi.12895
11. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403. https://doi.org/10.1055/s-0038-1634558
12. Helwig A. EHR certification criteria for SNOMED CT will help doctors transition to ICD-10 [Internet]. Washington (DC): Office of the National Coordinator for Health Information Technology (ONC); 2013 [cited at 2022 Jun 30]. Available from: https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/ehr-certification-criteria-snomed-ct-doctors-transition-icd10.
13. eMERGE Network. Epilepsy/antiepileptic drug response algorithm [Internet]. Nashville (TN): Vanderbilt University; c2017 [cited at 2022 Jun 30]. Available from: https://phekb.org/phenotype/epilepsyantiepileptic-drug-response-algorithm.


https://doi.org/10.4258/hir.2022.28.3.240