Original Article
Yonsei Med J 2025 Mar;66(3):187-194
https://doi.org/10.3349/ymj.2023.0628
Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Digital Phenotyping of Rare Endocrine Diseases Across International Data Networks and the Effect of Granularity of Original Vocabulary

Seunghyun Lee[1,2*], Namki Hong[1,3*], Gyu Seop Kim[3], Jing Li[4], Xiaoyu Lin[4], Sarah Seager[4], Sungjae Shin[1], Kyoung Jin Kim[5], Jae Hyun Bae[6,7], Seng Chan You[3,8], Yumie Rhee[1], and Sin Gon Kim[5]

[1]Department of Internal Medicine, Endocrine Research Institute, Yonsei University College of Medicine, Seoul, Korea;
[2]Department of Internal Medicine, Yonsei University Wonju College of Medicine, Wonju, Korea;
[3]Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Korea;
[4]Real-World Solutions, IQVIA, Durham, USA;
[5]Department of Internal Medicine, Korea University College of Medicine, Seoul, Korea;
[6]Department of Internal Medicine, Korea University Anam Hospital, Seoul, Korea;
[7]Department of Internal Medicine, Seoul National University Hospital, Seoul, Korea;
[8]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea.

Purpose: Rare diseases occur in <50 per 100000 people and require lifelong management. However, essential epidemiological data on such diseases are lacking, and a consecutive monitoring system across time and regions remains to be established. Standardized digital phenotypes are required to leverage an international data network for research on rare endocrine diseases. We developed digital phenotypes for rare endocrine diseases using the observational medical outcome partnership common data model.

Materials and Methods: Digital phenotypes of three rare endocrine diseases (medullary thyroid cancer, hypoparathyroidism, pheochromocytoma/paraganglioma) were validated across three databases that use different vocabularies: Severance Hospital's electronic health record from South Korea; IQVIA's United Kingdom (UK) database for general practitioners; and IQVIA's United States (US) hospital database for general hospitals. We estimated the performance of different digital phenotyping methods based on International Classification of Diseases (ICD)-10 in the UK and the US or systematized nomenclature of medicine clinical terms (SNOMED CT) in Korea.

Results: The positive predictive value of digital phenotyping was higher using SNOMED CT-based phenotyping than ICD-10-based phenotyping for all three diseases in Korea (e.g., pheochromocytoma/paraganglioma: ICD-10, 58%–62%; SNOMED CT, 89%). Estimated incidence rates by digital phenotyping were as follows: medullary thyroid cancer, 0.34–2.07 (Korea), 0.13–0.30 (US); hypoparathyroidism, 0.40–1.20 (Korea), 0.59–1.01 (US), 0.00–1.78 (UK); and pheochromocytoma/paraganglioma, 0.95–1.67 (Korea), 0.35–0.77 (US), 0.00–0.49 (UK).

Conclusion: Our findings demonstrate the feasibility of developing digital phenotyping of rare endocrine diseases and highlight the importance of implementing SNOMED CT in routine clinical practice to provide granularity for research.

Key Words: Common data model, digital phenotyping, rare diseases, medullary thyroid cancer, hypoparathyroidism, pheochromocytoma

Received: February 19, 2024 **Revised:** July 3, 2024 **Accepted:** August 12, 2024 **Published online:** November 28, 2024
Co-corresponding authors: Yumie Rhee, MD, PhD, Department of Internal Medicine, Endocrine Research Institute, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.
E-mail: YUMIE@yuhs.ac and
Seng Chan You, MD, PhD, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.
E-mail: CHANDRYOU@yuhs.ac

*Seunghyun Lee and Namki Hong contributed equally to this work.
•Seng Chan You is the Chief Technology Officer of PHI Digital Healthcare. The other authors have no potential conflicts of interest to disclose.

INTRODUCTION

Rare diseases, which occur in <50 per 100000 people, require continuous and often lifelong management, and consume substantial medical resources.[1] More than 6000 rare diseases affect approximately 30 million individuals across Europe and the United States (US).[2] Together, lack of awareness, data scarcity regarding the disease burden, progress and complications, and a limited number of specialists, prolong the diagnostic period by approximately 4 years, a phenomenon termed "diagnostic odyssey".[3] Due to the challenges in diagnosis and the lack of cost-effectiveness in research caused by their rarity, large investment in infrastructure for epidemiological studies of rare diseases seems remote.[4] Although more than 440 rare endocrine diseases exist across a wide range of organs including pituitary, thyroid, and adrenal glands, ovaries, testes, and bones, essential epidemiological data on these diseases remain lacking.[4,5] For example, the prevalence of nonsurgical hypoparathyroidism has been reported in approximately five countries, but no accurate incidence has been reported.[6]

Observational data from multiple healthcare resources, such as electronic health records (EHR), can provide evidence for rare diseases.[7,8] In addition, strengthening global research and collaboration by connecting medical specialists and their data offers opportunities to improve care for rare diseases.[9] However, the explosive growth of available data in healthcare has not supported a consecutive surveillance system for rare diseases across time and regions owing to inconsistencies in data format, semantic heterogeneity, and lack of consensus on digital phenotypes.[10] The common data model (CDM), which standardized data contained in different databases into a standardized format with a standardized vocabulary, can address this problem.[11] Converting multiple heterogeneous databases to CDM allows researchers to execute identical study protocols across databases with minimal effort.[12] The initial mapping of the local code to the standard concepts of the CDM in each database requires detailed knowledge of the local data; however, once the local code is translated into a common representation, the detailed knowledge requirement is minimal.[12] The observational medical outcome partnership (OMOP) CDM has a comprehensive vocabulary and scheme,[13,14] which are developed and maintained by the observational health data sciences and informatics (OHDSI) standardized vocabularies.[11,15] The OHDSI research network based on OMOP CDM is capable of building multinational and large-scale observational data networks worldwide.[16]

To leverage this international data network for research on rare endocrine diseases, it is crucial to develop and validate a standardized set of digital phenotypes. Digital phenotypes refer to a set of clinical concepts and pseudocode representations of clinical practices.[17] The effect of vocabulary mapping on fidelity and transportability in digital phenotypes has been rigorously validated across OMOP CDM databases for common diseases, including heart failure, diabetes mellitus, appendicitis, and cataracts.[18] Despite the comprehensive nature of the OMOP vocabulary for rare diseases, as demonstrated by Zoch, et al.,[9] the influence of vocabulary mapping and the granularity of the original vocabulary on digital phenotyping for rare diseases has yet to be explored.

In the present study, we developed and validated digital phenotypes for three rare endocrine diseases based on OMOP CDM: medullary thyroid cancer, nonsurgical hypoparathyroidism, and pheochromocytoma/paraganglioma. In addition, we evaluated the impact of granularity of vocabulary in the International Classification of Diseases (ICD)-10 systematized nomenclature of medicine clinical terms (SNOMED CT), and Read vocabulary.[19]

MATERIALS AND METHODS

Data sources

Data were obtained from three databases: rare endocrine diseases (RED)–CDM cohort from EHRs in Severance Hospital, one of the largest tertiary medical institutions in South Korea; IQVIA medical research database for UK general practitioners, known as the health improvement network (THIN); and IQVIA US hospital database for general hospitals.[20] The EHR system at Severance Hospital was constructed using both ICD-10 and SNOMED CT vocabularies, which we subsequently utilized in our research.

OMOP standard vocabulary

The healthcare systems of different countries usually have their own vocabulary for medical conditions. The US and Korea implement country-specific ICD-10 vocabularies: ICD-10-CM (US) and KCD7 (Korea), respectively. The UK has developed and maintained Read codes for the national health system.[21] In the OMOP CDM, the diverse original vocabularies for medical conditions are mapped into the OMOP standard vocabulary based on SNOMED CT.

Evaluation of sensitivity and positive predictive value

We extracted the entire patient cohort from the Severance Hospital data if they had at least one diagnosis of medullary thyroid cancer, hypoparathyroidism, or pheochromocytoma/paraganglioma. A through manual review was then conducted to evaluate the actual disease status. Following this, patients who were classified as part of the disease group according to each digital phenotyping were identified.

To estimate the sensitivity and positive predictive value (PPV), three specialists (Namki Hong, Kyoung Jin Kim, and Seunghyun Lee) reviewed the EHR data and confirmed the diagnosis. In case of disagreement, an expert with more than 10 years of experience (Yumie Rhee) made the final decision. PPV was calculated as the proportion of patients with diseases

https://doi.org/10.3349/ymj.2023.0628

confirmed by physicians among those detected using the digital phenotyping. Sensitivity was calculated as the proportion of patients with positive results based on the digital phenotype among those with an actual disease in the RED–CDM.

Study population and digital phenotyping

The digital phenotyping performance was confirmed by applying it to the RED–CDM cohort from January 1, 2011 to December 31, 2020 at Severance Hospital. Digital phenotyping based on ICD-10 and reported in prior literature was defined as ICD-10-originated OMOP concepts-level 1 (ICD-10-1).[22-24] If the PPV and sensitivity in ICD-10-1 did not reach at least 80%, the inclusion and exclusion criteria for each digital phenotyping were updated under discussion among experts, and this was defined as ICD-10-originated OMOP concepts-level 2 (ICD-10-2). In each disease, digital phenotyping was revised for the following reasons.

Medullary thyroid cancer

Experts' manual review suggested that the low sensitivity of ICD-10-1 in medullary thyroid cancer may be attributed to the exclusion criteria of a history of thyroglobulin test. Thyroglobulin test is a tumor marker to evaluate the therapeutic effect and monitor recurrence in patients with papillary and follicular thyroid cancer; this test is generally performed in patients with thyroid cancer.[25] In the RED–CDM cohort, 120 patients with medullary thyroid cancer who underwent thyroglobulin testing after surgery were observed. Therefore, in phenotyping based on ICD-10-2, the exclusion condition of a history of thyroglobulin test was deleted.

Nonsurgical hypoparathyroidism

The PPV of ICD-10-1 was lower than expected compared to the PPV of 91% reported in previous studies.[26,27] Due to the nature of CDM, confirming whether surgery was performed at another hospital, along with diagnosis and laboratory test codes was challenging. Most cases of surgical hypoparathyroidism that occurred after thyroid surgery at other hospitals were erroneously classified as nonsurgical hypoparathyroidism at our institution. To correct this misclassification, we excluded patients who were taking levothyroxine.

Pheochromocytoma/paraganglioma

The low PPV of ICD-10-1 was attributed to issues with the definition of catecholamine measurement—a total of two or more catecholamine tests, including one or more tests before surgery. Until the 2023 guideline update, adrenal incidentalomas were monitored annually with hormonal assessments, including catecholamine measurements, for a period of 4 to 5 years.[28,29] Surgical intervention was indicated if an increase in size or abnormal adrenal function was detected during the follow-up.[28] Consequently, under the ICD-10-1 definition, not only patients diagnosed with pheochromocytoma but also those who underwent surgery after more than 2 years of hormonal follow-up testing, despite not having pheochromocytoma, were erroneously classified as part of the disease group. Therefore, for phenotyping based on ICD-10-2, the inclusion criteria of catecholamine measurements were revised to require at least one catecholamine test before surgery and more than one catecholamine test after surgery. Meanwhile, neuroblastoma arises from primitive sympathetic ganglion cells and can secrete catecholamines; in some cases, neuroblastoma is indistinguishable from pheochromocytoma/paraganglioma based on preoperative clinical features alone. Therefore, a history of being diagnosed with neuroblastoma at least once during the study period was added to the exclusion criteria.

In addition, when the SNOMED CT code was applied instead of the ICD-10 code as a cohort entry condition in the ICD-10-2 definition, it was defined as SNOMED CT-originated OMOP concepts (Fig. 1). The difference between each entry condition code in the ICD-10 and SNOMED CT used in digital phenotyping is described in Supplementary Table 1 (only online). For example, in medullary thyroid cancer, ICD used the entry condition code of the primary malignancy of the thyroid gland (ICD-10: C73, concept code: 94098005), whereas SNOMED CT used a more detailed entry condition code for medullary thyroid carcinoma (concept code: 255032005). Based on this digital phe-

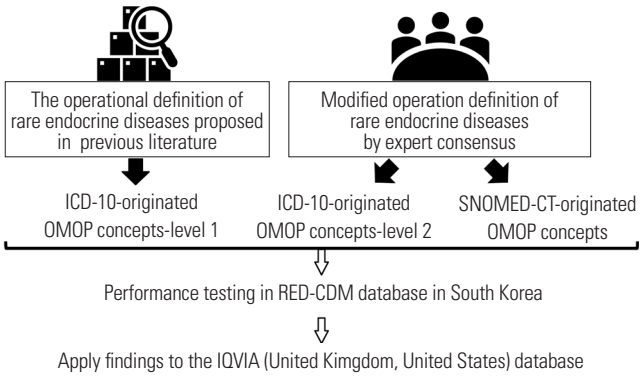

Fig. 1. Studyflow. In ICD-10-originated OMOP concepts-level 1, we used the operational definitions reported in previous Korean studies based on ICD-10. In ICD-10-originated OMOP concepts-level 2, we aimed to improve sensitivity or positive predictive value through expert consensus. The changes made from ICD-10-originated OMOP concepts-level 1 to ICD-10-originated OMOP concepts-level 2 were as follows: 1) for medullary thyroid cancer, we removed the exclusion criterion of "thyroglobulin test"; 2) for nonsurgical hypoparathyroidism, we added "patients taking levothyroxine" to the inclusion criteria; 3) for pheochromocytoma/paraganglioma, we revised the inclusion criterion from "measuring catecholamine at least twice during the entire period" to "measuring catecholamine at least once before and once after surgery". In SNOMED CT-originated OMOP concepts, the vocabulary was changed from ICD-10 to SNOMED CT. The performance of these three digital phenotyping was tested using the RED–CDM database in South Korea, and subsequently applied to the IQVIA database (United Kingdom and United States). ICD-10, International Classification of Diseases-10; OMOP, observational medical outcome partnership; SNOMED CT, systematized nomenclature of medicine clinical terms; RED–CDM, rare endocrine disease–common data model.

notyping, the number of patients at Severance Hospital and IQVIA (US and UK) was investigated (Table 1, Supplementary Figs. 1 and 2, only online).

Statistical analysis

The crude incidence rate was estimated using the total number of events divided by the total follow-up period duration for all patients. The incidence per 100000 person-years was calculated as the total number of events divided by the total number of person-years. Age was classified into nine age ranges: 0–9, 10–19, 20–29, 30-39, 40–49, 50–59, 60–69, 70–79, and 80–89 years. Patients older than 90 years were not included. All statistical analyses were performed using R version 4.1.3 (R Foundation for Statistical Computing, Vienna, Austria). The study protocol and analysis codes are available in the reference provided.[30]

RESULTS

Digital phenotyping verification

Each digital phenotyping was defined as described in Supplementary Fig. 2 (only online). The actual number of patients in the RED-CDM database for medullary thyroid cancer, nonsurgical hypoparathyroidism, and pheochromocytoma/paraganglioma were 130, 116, and 172, respectively. For medullary thyroid cancer, only 17 patients were identified using ICD-10-1 during the study period. The PPV and sensitivity of phenotyping using ICD-10-1 were 100% and 13%, respectively. Using ICD-10-2, the sensitivity increased from 13% to 100%, whereas PPV decreased to 96%. When using the SNOMED CT-originated OMOP concept, the sensitivity and PPV were 97% and 100%, respectively.

In nonsurgical hypoparathyroidism, the PPV of phenotyping based on ICD-10-1 was 72% and the sensitivity was 100%. After applying ICD-10-2, the PPV of ICD-10-2 phenotyping increased from 72% to 82%. When using SNOMED CT-originated OMOP concepts, the PPV improved from 82% to 84%.

In pheochromocytoma/paraganglioma, the PPV of ICD-10-1 was 58%, and the sensitivity was 90%. After applying ICD-10-2, the PPV increased from 58% to 62%. Moreover, in SNOMED CT-originated OMOP concepts, the PPV improved from 62% to 89% (Table 2).

Table 1. Characteristics of Included Populations in Each Database

Characteristics	RED-CDM_KR	IQVIA_US	IQVIA_UK
Total No. of patients	5.8 M	106.5 M	13.7 M
Person-years	13267073	96628400	8358871
Data type	EHR	EHR	EHR
Dates of service	2005–2022	2007–2022	1994–2022
Dates of research	2011–2020	2011–2020	2011–2020
Care sites	Single center	1.1 K	832
Age group (yr)			
0–9	432207 (13.5)	11209306 (16.7)	670282 (27.7)
10–19	219826 (6.9)	6667446 (9.9)	192940 (8.0)
20–29	593756 (18.6)	9032804 (13.4)	450238 (18.6)
30–39	472993 (14.8)	8559708 (12.7)	383195 (15.8)
40–49	429024 (13.4)	8092993 (12.0)	247147 (10.2)
50–59	449025 (14.0)	8825658 (13.1)	185478 (7.7)
60–69	336422 (10.5)	7840416 (11.7)	140031 (5.8)
70–79	198203 (6.2)	5902123 (8.8)	88466 (3.7)
80–89	43182 (1.4)	1095835 (1.6)	66382 (2.7)
Sex			
Male	1499624 (46.9)	30149373 (44.9)	1160409 (47.9)
Female	1698985 (53.1)	37076916 (55.2)	1263750 (52.1)

RED-CDM, rare endocrine disease-common data model; KR, South Korea; US, United States; UK, United Kingdom; No., number; M, million; EHR, electronic health records; K, thousand.
Data are presented as n (%).

Digital phenotyping applied to multicenter, international research of three rare endocrine diseases

We applied this digital phenotyping to IQVIA (US and UK) database. Detailed incidence of each disease is described in Supplementary Table 2 (only online).

Medullary thyroid cancer

In the RED–CDM for South Korea, the incidence rates per 100000 person-years defined by phenotyping based on ICD-10-1, ICD-10-2, and SNOMED CT were 0.09, 0.65, and 0.64, respectively. The crude incidence rates in South Korea for ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 0.38, 2.69, and 2.66, respectively. In the US cohort, the incidence rates per 100000 person-years of ICD-10-1 and ICD-10-2 phenotype were 0.06 and 0.08, respectively. The crude incidence rates in the US for ICD-10-1 and ICD-10-2 phenotype were 0.015 and 0.19, respectively. Meanwhile, the incidence using the SNOMED CT-originated OMOP concept in the US was 0. In the UK cohort, the incidence was confirmed as a zero count for all digital phenotyping.

Nonsurgical hypoparathyroidism

The incidence rates per 100000 person-years in the RED–CDM for South Korea defined by phenotyping based on ICD-10-1, ICD-10-2, and SNOMED CT were 0.77, 0.56, and 0.50, respectively. The crude incidence rates per 100000 persons of ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 3.19, 2.31, and 2.06, respectively. In the US cohort, the incidence rates per 100000 person-years for ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 0.76, 0.42, and 0.31, respectively. The crude incidence rates per 100000 persons in the US for ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 1.88, 1.04, and 0.77, respectively. In the UK cohort, the incidence rates per 100000 person-years for ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 0.07, 0.04, and 0.04, respectively. The crude incidence rates per 100000 persons in the UK for ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 1.28, 0.74, and 0.70,

https://doi.org/10.3349/ymj.2023.0628

Table 2. PPV and Sensitivity According to Each Digital Phenotyping Method in Korea

Cohort definitions	Number of patients detected by digital phenotyping (n)	Number of true positive patients (n)	Total number of actual patients in the cohort (n)	PPV (%)	Sensitivity (%)
Medullary thyroid cancer					
ICD-10-1	17	17	130	100	13
ICD-10-2	135	130	130	96	100
SNOMED CT	126	126	130	100	97
Nonsurgical hypoparathyroidism					
ICD-10-1	162	116	116	72	100
ICD-10-2	98	80	116	82	69
SNOMED CT	93	78	116	84	67
Pheochromocytoma/ paraganglioma					
ICD-10-1	266	155	172	58	90
ICD-10-2	231	144	172	62	84
SNOMED CT	159	141	172	89	82

ICD-10-1, ICD-10-originated OMOP concept-level 1; ICD-10-2, ICD-10-originated OMOP concept-level 2; PPV, positive predictive value; ICD-10, International Classification of Diseases-10; OMOP, observational medical outcome partnership; SNOMED CT, systematized nomenclature of medicine clinical terms.

Table 3. The Incidence Rate Per 100000 Person-Years of Medullary Thyroid Cancer, Nonsurgical Hypoparathyroidism, and Pheochromocytoma/Paraganglioma According to Digital Phenotyping

	Korea	US	UK
Medullary thyroid cancer			
ICD-10-1	0.09	0.06	N/A
ICD-10-2	0.65	0.08	N/A
SNOMED CT	0.64	N/A	N/A
Nonsurgical hypoparathyroidism			
ICD-10-1	0.77	0.76	0.07
ICD-10-2	0.56	0.42	0.04
SNOMED CT	0.50	0.31	0.04
Pheochromocytoma/paraganglioma			
ICD-10-1	1.45	0.42	0.02
ICD-10-2	1.28	0.23	0.01
SNOMED CT	0.87	N/A	0.01

ICD-10-1, ICD-10-originated OMOP concept level 1; ICD-10-2, ICD-10-originated OMOP concept level 2; KR, South Korea; US, United States; UK, United Kingdom; ICD-10, International Classification of Diseases-10; OMOP, observational medical outcome partnership; SNOMED CT, systematized nomenclature of medicine clinical terms

respectively.

Pheochromocytoma/paraganglioma

In the RED–CDM cohort of South Korea, the incidence rates per 100000 person-years for ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 1.45, 1.28, and 0.87, respectively. The crude incidence rates of ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 6.00, 5.31, and 3.63, respectively. In the US cohort, the incidence rate per 100000 person-years of ICD-10-1 phenotype was 0.42, and that of ICD-10-2 phenotype was 0.23. The crude incidence rate in the US for ICD-10-1 phenotype was 1.04, and that of ICD-10-2 phenotype was 0.57. The incidence using the SNOMED CT OMOP concept in the US was 0. In the UK cohort, the incidence rates per 100000 person-years for phenotyping based on ICD-10-1, ICD-10-2, and SNOMED CT were 0.02, 0.01, and 0.01, respectively. The crude incidence rates in the UK for ICD-10-1, ICD-10-2, and SNOMED CT phenotype were 0.33, 0.17, and 0.17, respectively (Table 3, Supplementary Fig. 3, only online).

Challenges in applying digital phenotyping to multicenter, international research

In the UK, zero counts were obtained from the thyroid cancer entry code (ICD-10 code C73) in the ICD-10-based definition of medullary thyroid cancer (Supplementary Table 3, only online). Failure in detection of the diagnostic codes was likely. Also, no SNOMED CT mapping was identified in the US corresponding to medullary thyroid cancer or pheochromocytoma/paraganglioma.

DISCUSSION

In this study, the digital phenotyping was developed for three rare endocrine diseases: medullary thyroid cancer, nonsurgical hypoparathyroidism, and pheochromocytoma/paraganglioma. The performance of digital phenotyping was investigated and indicated that using SNOMED CT with high vocabulary granularity can be helpful regarding rare endocrine diseases. We also proposed ICD-10-originated concepts of three diseases, referred to as ICD-10-2, which demonstrated comparable sensitivity and PPV to SNOMED CT-originated OMOP concepts.

The main finding of our study, that adopting SNOMED CT in routine clinical practice provides improved granularity in rare endocrine diseases, is consistent with previous literature. Several studies have compared the coverage between different vocabularies. For each concept, including diagnoses, treatments, and procedures, SNOMED CT showed greater coverage compared to ICD-10 and ICD-9-CM. In addition,

studies conducted on rare diseases indicated that SNOMED CT could support research and evidence-based care due to its improved coverage.[31,32] A study of 6519 rare diseases using the Unified Medical Language System showed that 11% of disorders matched ICD-9-CM, 21% matched ICD-10, and 44% matched SNOMED CT.[32] Furthermore, ICD-10 coding for rare diseases is insufficient; only 300 of 6000–8000 diseases can be documented using ICD-10.[33] A previous study using the Danish national registry also demonstrated that the detailed granularity of SNOMED CT aids in the identification of pheochromocytoma better than ICD-8 or ICD-10.[34]

However, compared to previous studies, our study had some limitations due to the lack of epidemiological data, with only one study reporting age-standardized incidence rates and no study reporting crude incidence rates (Supplementary Table 4, only online). This study revealed a certain tendency when compared with previous research; in South Korea, the incidence using phenotyping based on ICD-10-1 differed from that reported in previous nationwide cohort studies.[22-24] In contrast, ICD-10-2 and SNOMED CT phenotyping showed similar or higher incidence rates to those previously reported. Considering that the RED–CDM uses data from tertiary referral centers, a higher incidence rate in the RED-CDM cohort than those reported in previous studies is reasonable. In particular, the incidence rates of phenotyping based on ICD-10-1 and ICD-10-2 were higher than those reported at the Mayo Clinic in the US, using RED–CDM, whereas the incidence rate of phenotyping based on SNOMED CT was similar to that reported at the Mayo Clinic.[35] This finding suggests that SNOMED CT could provide accurate results. Also, the incidence in this study was similar to that reported in the US[36] and lower than that reported in the UK.[37,38] The difference between the incidence rates reported in previous literature from the UK and our study may be related to the fact that the UK database mainly includes primary care, whereas previous studies were based on the hospital setting.[37] In both the US and UK, zero counts were observed for certain rare endocrine diseases due to a mismatch of diagnostic codes in our study. The reason for the zero count results of SNOMED CT-based digital phenotyping in the US data may be due to the insufficient granularity of the diagnostic inputs. In the UK THIN database, the reason for the absence of all digital phenotyping of medullary thyroid cancer is unclear; however, it may be influenced by mismatched or unmatched diagnostic codes. This finding is consistent with previous literature on UK Biobank codes, which demonstrated that only 2.7% of Read codes successfully mapped to ICD-10-codes.[39] These results highlight the challenges of conducting international epidemiological studies using standardized coding systems like ICD-10, especially for rare disease, as they may present obstacles to consistent implementation across diverse settings.

While the OMOP CDM provides a comprehensive ontology system to harmonize international data semantically,[15] there are limitations in that most data sources are routinely collected; and the analysis is the secondary use of those data. As previously shown by Ostropolets, et al.,[40] heterogeneity in the granularity level across data sources can affect the performance of digital phenotyping.

To the best of our knowledge, this is the first study to present and validate digital phenotypes of rare endocrine diseases in multinational databases. This study may serve as a reference for future researchers who wish to undertake similar data mapping projects. In addition, our study suggests that the global monitoring of rare diseases is possible through CDM. Moreover, we demonstrated the strength of adopting SNOMED CT in routine clinical practice for monitoring rare endocrine diseases. These findings have important implications for global health policy regarding rare disease management. Policymakers should consider mandating or incentivizing the adoption of granular terminologies like SNOMED CT in healthcare systems to enable more effective monitoring and research of rare diseases. Such policies could facilitate the development of international collaborations and data sharing initiatives, ultimately leading to improved understanding, diagnosis, and treatment of rare diseases worldwide.

A limitation of this study is that we did not investigate vocabularies other than ICD-10, SNOMED CT, and Read codes. Phenotyping for another rare disease may be more accurate when using terminology from other specialty fields, including the NCI Thesaurus for cancer description[41] and RadLex for radiological text.[42] Moreover, the superior granularity of SNOMED CT compared to ICD-10 definitions is particularly evident in one of the three diseases in our study; however, when classifying the presence of diseases based solely on diagnostic codes, without using digital phenotyping with operational definitions, SNOMED CT demonstrated greater accuracy compared to ICD-10 (data not shown). Also, comparing the incidence rates of this rare endocrine diseases across different countries has been challenging due to the lack of reported data. Future research should expand the scope of this study to include a broader range of rare diseases, leveraging specialty-specific terminologies and data to enhance phenotype accuracy by integrating genomic,[43] imaging,[44] and text data.[45] Additionally, future studies should aim to include a larger number of data sources from diverse healthcare settings and geographic regions to further validate the generalizability of the digital phenotypes and assess the impact of different data capturing systems and coding practices on phenotype performance.

In summary, our study developed the digital phenotyping of three rare endocrine diseases, which may prove useful in future studies related to rare endocrine diseases. Furthermore, we showed that rare endocrine diseases can benefit from using SNOMED CT with a high vocabulary granularity. We also demonstrated that by applying various concepts, ICD-10 digital phenotyping can achieve performance comparable to SNOMED CT, making it applicable even in cases where ICD-10 is required.

https://doi.org/10.3349/ymj.2023.0628

DATA STATEMENT

Since this was a distributed data analysis, individual patient-level data from each database cannot be shared owing to database governance restrictions. The prespecified protocol and executable source code are available online (https://github.com/redcdm/REDCDM_RedCohort).

ACKNOWLEDGEMENTS

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C0189).

AUTHOR CONTRIBUTIONS

Conceptualization: Namki Hong, Jae Hyun Bae, Seng Chan You, Yumie Rhee, and Sin Gon Kim. Data curation: Seunghyun Lee, Namki Hong, Gyu Seop Kim, Jing Li, Xiaoyu Lin, Sungjae Shin, Kyoung Jin Kim, and Jae Hyun Bae. Formal analysis: Seunghyun Lee, Namki Hong, and Gyu Seop Kim. Funding acquisition: Namki Hong, Yumie Rhee, and Sin Gon Kim. Investigation: Seunghyun Lee, Namki Hong, Gyu Seop Kim, Jing Li, Xiaoyu Lin, and Sarah Seager. Methodology: Namki Hong and Seng Chan You. Project administration: Namki Hong, Seng Chan You, and Yumie Rhee. Resources: Namki Hong, Sungjae Shin, Kyoung Jin Kim, Seng Chan You, Yumie Rhee, and Sin Gon Kim. Supervision: Namki Hong, Seng Chan You, and Yumie Rhee. Visualization: Seunghyun Lee, Namki Hong, and Gyu Seop Kim. Writing—original draft: Seunghyun Lee and Namki Hong. Writing—review & editing: Seng Chan You and Yumie Rhee. Approval of final manuscript: all authors.

ORCID iDs

Seunghyun Lee https://orcid.org/0000-0001-9254-183X
Namki Hong https://orcid.org/0000-0002-8246-1956
Gyu Seop Kim https://orcid.org/0009-0004-6014-4177
Jing Li https://orcid.org/0009-0008-8187-0951
Xiaoyu Lin https://orcid.org/0009-0003-4292-9547
Sarah Seager https://orcid.org/0000-0002-0614-4166
Sungjae Shin https://orcid.org/0000-0003-3213-580X
Kyoung Jin Kim https://orcid.org/0000-0001-7925-2515
Jae Hyun Bae https://orcid.org/0000-0002-1384-6123
Seng Chan You https://orcid.org/0000-0002-5052-6399
Yumie Rhee https://orcid.org/0000-0003-4227-5638
Sin Gon Kim https://orcid.org/0000-0002-7430-3675

REFERENCES

1. Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, et al. Rare disease terminology and definitions—a systematic global review: report of the ISPOR Rare Disease Special Interest Group. Value Health 2015;18:906-14.
2. The Lancet Diabetes & Endocrinology. Spotlight on rare diseases. Lancet Diabetes Endocrinol 2019;7:75.
3. Hartley T, Lemire G, Kernohan KD, Howley HE, Adams DR, Boycott KM. New diagnostic approaches for undiagnosed rare genetic diseases. Annu Rev Genomics Hum Genet 2020;21:351-72.
4. Reincke M, Hokken-Koelega A. Perspectives of the European Society of Endocrinology (ESE) and the European Society of Paediatric Endocrinology (ESPE) on rare endocrine disease. Endocrine 2021;71:539-41.
5. Marcucci G, Cianferotti L, Beck-Peccoz P, Capezzone M, Cetani F, Colao A, et al. Rare diseases in clinical endocrinology: a taxonomic classification system. J Endocrinol Invest 2015;38:193-259.
6. Clarke BL. Epidemiology and complications of hypoparathyroidism. Endocrinol Metab Clin North Am 2018;47:771-82.
7. Cheng HG, Phillips MR. Secondary analysis of existing data: opportunities and implementation. Shanghai Arch Psychiatry 2014;26:371-5.
8. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395-405.
9. Zoch M, Gierschner C, Peng Y, Gruhl M, Leutner LA, Sedlmayr M, et al. Adaption of the OMOP CDM for rare diseases. Stud Health Technol Inform 2021;281:138-42.
10. Forrest CB, Bartek RJ, Rubinstein Y, Groft SC. The case for a global rare-diseases registry. Lancet 2011;377:1057-9.
11. 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.
12. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19:54-60.
13. 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.
14. Liyanage H, Liaw ST, Jonnagaddala J, Hinton W, de Lusignan S. Common data models (CDMs) to enhance international big data analytics: a diabetes use case to compare three CDMs. Stud Health Technol Inform 2018;255:60-4.
15. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI standardized vocabularies-a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc 2024;31:583-90.
16. You SC, Lee S, Choi B, Park RW. Establishment of an international evidence sharing network through common data model for cardiovascular research. Korean Circ J 2022;52:853-64.
17. 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.
18. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. J Am Med Inform Assoc 2018;25:1618-25.
19. Berger A, Rustemeier AK, Göbel J, Kadioglu D, Britz V, Schubert K, et al. How to design a registry for undiagnosed patients in the framework of rare disease diagnosis: suggestions on software, data set and coding system. Orphanet J Rare Dis 2021;16:198.
20. The Health Improvement Network. Main page [Internet] [accessed on 2022 August 25]. Available at: https://www.the-health-improvement-network.com.
21. Gray J, Orr D, Majeed A. Use of Read codes in diabetes management in a south London primary care group: implications for establishing disease registers. BMJ 2003;326:1130.
22. Ahn HY, Chae JE, Moon H, Noh J, Park YJ, Kim SG. Trends in the diagnosis and treatment of patients with medullary thyroid carcinoma in Korea. Endocrinol Metab (Seoul) 2020;35:811-9.
23. Kim JH, Moon H, Noh J, Lee J, Kim SG. Epidemiology and prognosis of pheochromocytoma/paraganglioma in Korea: a nationwide study based on the National Health Insurance Service. En-

docrinol Metab (Seoul) 2020;35:157-64.
24. Kim SH, Rhee Y, Kim YM, Won YJ, Noh J, Moon H, et al. Prevalence and complications of nonsurgical hypoparathyroidism in Korea: a nationwide cohort study. PLoS One 2020;15:e0232842.
25. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016;26:1-133.
26. Swartling O, Evans M, Spelman T, Kamal W, Kämpe O, Mannstadt M, et al. Kidney complications and hospitalization in patients with chronic hypoparathyroidism: a cohort study in Sweden. J Clin Endocrinol Metab 2022;107:e4098-105.
27. Kamal W, Björnsdottir S, Kämpe O, Trolle Lagerros Y. Concordance between ICD-10 codes and clinical diagnosis of hypoparathyroidism in Sweden. Clin Epidemiol 2020;12:327-31.
28. Lee JM, Kim MK, Ko SH, Koh JM, Kim BY, Kim SW, et al. Clinical guidelines for the management of adrenal incidentaloma. Endocrinol Metab (Seoul) 2017;32:200-18.
29. Fassnacht M, Tsagarakis S, Terzolo M, Tabarin A, Sahdev A, Newell-Price J, et al. European Society of Endocrinology clinical practice guidelines on the management of adrenal incidentalomas, in collaboration with the European Network for the Study of Adrenal Tumors. Eur J Endocrinol 2023;189:G1-42.
30. Hong N, Kim K, Bae JH, You SH, Park Y, Kim KJ, et al. Rare endocrine disease common data model (RED-CDM) [accessed on 2022 August 25]. Available at: https://github.com/QSB-yuhs/REDCDM_RedCohort/blob/master/extras/REDCDM_OHDSI_Protocol.docx.
31. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. J Am Med Inform Assoc 1996;3:224-33.
32. Fung KW, Richesson R, Bodenreider O. Coverage of rare disease names in standard terminologies and implications for patients, providers, and research. AMIA Annu Symp Proc 2014;2014:564-72.
33. Aymé S, Bellet B, Rath A. Rare diseases in ICD11: making rare diseases visible in health information systems through appropriate coding. Orphanet J Rare Dis 2015;10:35.
34. Ebbehoj A, Jacobsen SF, Trolle C, Robaczyk MG, Rasmussen ÅK, Feldt-Rasmussen U, et al. Pheochromocytoma in Denmark during 1977-2016: validating diagnosis codes and creating a national cohort using patterns of health registrations. Clin Epidemiol 2018;10:683-95.
35. Beard CM, Sheps SG, Kurland LT, Carney JA, Lie JT. Occurrence of pheochromocytoma in Rochester, Minnesota, 1950 through 1979. Mayo Clin Proc 1983;58:802-4.
36. Randle RW, Balentine CJ, Leverson GE, Havlena JA, Sippel RS, Schneider DF, et al. Trends in the presentation, treatment, and survival of patients with medullary thyroid cancer over the past 30 years. Surgery 2017;161:137-46.
37. Cvasciuc IT, Gull S, Oprean R, Lim KH, Eatock F. Changing pattern of pheochromocytoma and paraganglioma in a stable UK population. Acta Endocrinol (Buchar) 2020;16:78-85.
38. Vadiveloo T, Donnan PT, Leese GP. A population-based study of the epidemiology of chronic hypoparathyroidism. J Bone Miner Res 2018;33:478-85.
39. Stroganov O, Fedarovich A, Wong E, Skovpen Y, Pakhomova E, Grishagin I, et al. Mapping of UK Biobank clinical codes: challenges and possible solutions. PLoS One 2022;17:e0275816.
40. Ostropolets A, Reich C, Ryan P, Weng C, Molinaro A, DeFalco F, et al. Characterizing database granularity using SNOMED-CT hierarchy. AMIA Annu Symp Proc 2021;2020:983-92.
41. Schulz S, Daumke P, Romacker M, López-García P. Representing oncology in datasets: standard or custom biomedical terminology? Inform Med Unlocked 2019;15:100186.
42. Kahn CE Jr. Annotation of figures from the biomedical imaging literature: a comparative analysis of RadLex and other standardized vocabularies. Acad Radiol 2014;21:384-92.
43. Shin SJ, You SC, Park YR, Roh J, Kim JH, Haam S, et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. J Med Internet Res 2019;21:e13249.
44. Park WY, Jeon K, Schmidt TS, Kondylakis H, Alkasab T, Dewey BE, et al. Development of medical imaging data standardization for imaging-based observational research: OMOP common data model extension. J Imaging Inform Med 2024;37:899-908.
45. Park J, You SC, Jeong E, Weng C, Park D, Roh J, et al. A framework (SOCRATex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study. JMIR Med Inform 2021;9:e23983.

https://doi.org/10.3349/ymj.2023.0628