ICD2Vec: Mathematical representation of diseases

Yeong Chan Lee

Samsung Advanced Institute for Health Science and Technology, Sungkyunkwan University School of Medicine <https://orcid.org/0000-0002-2093-3161>

Sang-Hyuk Jung

Sungkyunkwan University

Aman Kumar

Indian Institute of Technology Kharagpur

Injeong Shim

Sungkyunkwan University

Minku Song

Sungkyunkwan University

Min Seo Kim

Sungkyunkwan University

Kyunga Kim

Samsung Medical Center

Woojae Myung

Seoul National University Bundang Hospital

Woong-Yang Park

Samsung Medical Center

Hong-Hee Won (✉ honghee.won@gmail.com)

Sungkyunkwan University

Article

Keywords: International Classification of Diseases (ICD), global disease reporting, ICD2Vec, ICD codes

Posted Date: July 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-692012/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

ICD2Vec: Mathematical representation of diseases

Yeong Chan Lee^{1,&}, Sang-Hyuk Jung^{1,2,&}, Aman Kumar^{3,&}, Injeong Shim¹, Minku Song¹, Min Seo Kim¹, Kyunga Kim^{1,4}, Woojae Myung⁵, Woong-Yang Park⁶, Hong-Hee Won^{1,6*}

¹Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Samsung Medical Center, Seoul 06355, Republic of Korea

²Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia 19104, USA

³Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, West Bengal 721302, India

⁴Statistics and Data Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul 06355, Republic of Korea

⁵Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam 13620, Republic of Korea

⁶Samsung Genome Institute, Samsung Medical Center, Seoul 06351, Republic of Korea

[&]These authors contributed equally

*Corresponding author, wonhh@skku.edu

Abstract

The International Classification of Diseases (ICD) codes represent the global standard for reporting disease conditions. The current ICD codes are hierarchically structured and only connote partial relationships among diseases. Therefore, it is important to represent the ICD codes as mathematical vectors to indicate the complex relationships across diseases. Here, we proposed a framework denoted “ICD2Vec” for providing mathematical representations of diseases by encoding corresponding information. First, we presented the arithmetic and semantic relationships between diseases by mapping composite vectors for symptoms or diseases to the most similar ICD codes. Second, we confirmed the validity of ICD2Vec by comparing the biological relationships and cosine similarities among the vectorized ICD codes. Third, we proposed a new risk score derived from ICD2Vec, and demonstrated its potential clinical utility for coronary artery disease, type 2 diabetes, dementia, and liver cancer, based on a large prospective cohort from the UK and large electronic medical records from a medical centre in South Korea. In summary, ICD2Vec is applicable for diverse quantitative analyses using ICD codes in biomedical research.

In the past few decades, hospitals worldwide have rapidly adopted electronic health records (EHRs). Recently, more than 84% and 94% of hospitals have adopted EHRs in the UK and US, respectively^{1,2}. Standardized coding systems are the basis for rapid advances in healthcare systems. The tenth revision of the International Classification of Diseases (ICD-10) is a global standard for health data, clinical documentation, and statistical aggregation^{3,4}. It defines categorical codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injuries or diseases⁵⁻⁷. These diagnosis codes are practical because they provide a common ground for the reporting and monitoring of diseases, allowing hospitals to share data in a consistent way. However, despite the complex relationships across diseases, current ICD codes only contain hierarchical classifications⁴.

In addition to their importance in clinical practice, ICD codes are essential for biomedical research, e.g., to define the progress in disease conditions and as predictor and outcome variables for analyses of prospective disease risks⁸⁻¹². In epidemiology, the variables for the prevalence of a disease are conventionally coded using a binary dummy coding representing the presence or absence of the disease, regardless of the implicit relationships between the predictor and outcome variables^{13,14}. In machine learning, the widely used one-hot encoding can also miss such relationships between variables, leading to a decrease in performance or slow convergence speed. In contrast, if the mathematically defined ICD variables contain the relational features between diseases, the performances of corresponding models may be better than simple one-hot encoding-based models. Efficient embedding is important for reducing the dimensionality of the input data while maintaining the intrinsic information, thus enabling the development of a lightweight model¹⁵. The contextual embedding of clinical concepts has been investigated in previous studies (**Supplementary Information**)^{16,17}.

Nonetheless, there are no publicly available generalized vectors for providing mathematical representations of ICD codes that are universally applicable in biomedical research. Hence, an objective representation of ICD codes as mathematical vectors is necessary.

In this work, we proposed ICD2Vec, a framework for converting ICD codes into meaningful vectors (**Fig. 1**). In our study, we used word embeddings generated from FastText to construct a vector corresponding to each ICD code, such that each vector encoded meaningful information through the semantic relationships between the ICD codes. We demonstrated that the semantic relationships were preserved between ICD codes by finding an ICD code acting as an arithmetical combination of multiple symptoms or ICD codes. In addition, to confirm the validity of the vectorized ICD codes, we compared the semantic relationships with the biological relationships between diseases generated from other sources¹⁸. Finally, to illustrate the clinical utility of ICD2Vec, we proposed an individual risk score derived from ICD2Vec for predicting the risk of subsequent diseases based on a history of diagnosed diseases of an individual relative to a baseline. The score was applied to epidemiological studies using two datasets (from the UK and South Korea).

Fig. 1 | ICD2Vec overview. **a**, Illustration for encoding International Classification of Diseases (ICD) codes to vectors. We obtained disease definitions, clinical information, and approximate synonyms for all 19,154 ICD codes from <https://www.icd10data.com>. After natural language processing, every word was converted to

a corresponding 300-dimensional vector. Finally, we calculated an average for the word vectors to indicate an ICD code. **b.** We validated the ICD2Vec with analogical reasoning using arithmetic operations and observed significant associations in the cosine similarities between pairs of ICD codes and their biological relationships. We suggested a new “ICD2Vec risk score” (IRIS) derived from ICD2Vec. The IRIS is a value representing the cosine similarities between the target disease and diagnosed diseases for a person in a specific time period. The individual risk for the target disease was measured using IRIS according to the diagnosis history. We validated clinical utility of IRIS as a risk score with two datasets from the UK and South Korea. Illustrations are created with BioRender.com.

Results

We embedded the information regarding diseases from the ICD codes into vectors using the pretrained FastText model; this model was trained to learn the relationships between adjacent vectors from a large corpus and subword information.

Arithmetic and semantic relationships between ICD2Vec

We empirically investigated the relationships between a symptom and a disease or between diseases using the analogical reasoning task shown by Mikolov *et al.*^{19,20}. The task was performed by finding and evaluating the ICD code closest to the composite vector derived from the given vectors for the symptoms or diseases (**Table 1**). For example, the ICD code most similar to nearsightedness appears to be strabismus (ICD-10: H50); nearsightedness is one of the symptoms of strabismus^{21,22}. A specific symptom could be represented by arithmetic operations, such as addition and subtraction. For instance, itchy skin is denoted by the addition of the vector for “skin” and the vector for “itching,” and the composite vector derived from these two vectors matches with pruritus (ICD-10: L29). Using the same approach, it is possible to use the calculation of diseases (e.g., I25 + I10) to find clinical manifestations with similar medical descriptions (e.g., I51). We created a vector for coronavirus disease (COVID-19), as summarized by the Mayo clinic²³. Interestingly, the disease most similar to COVID-19 is A99, an unspecified viral hemorrhagic fever.

Table 1 | Examples of analogical reasoning with ICD2Vec. The top three similar International Classification of Diseases (ICD)-10 codes were found based on the cosine similarity with a given vector for a symptom or disease. Semantic compositionality through vector operations such as addition and subtraction was also possible when representing an ICD-10 code. Python codes are provided to check for the most similar ICD-10 code for the given symptoms or diseases (see Code availability).

Symptoms	Top three similar ICD-10 (definition of the disease) (CS)
Nearsightedness	H50 (Other strabismus) (0.61), H52 (Disorders of refraction and accommodation) (0.60), H57 (Other disorders of eye and adnexa) (0.58)
Skin + itching	L29 (Pruritus) (0.79), R23 (Other skin changes) (0.77), C44 (Other and unspecified malignant neoplasm of skin) (0.77)
Senseless + unconsciousness	R40 (Somnolence, stupor and coma) (0.75), R45 (Symptoms and signs involving emotional state) (0.70), F44 (Dissociative and conversion disorders) (0.69)
Breathing + problems	R06 (Abnormalities of breathing) (0.76), K30 (Functional dyspepsia) (0.70), R11 (Nausea and vomiting) (0.70)
Stomach + Neoplasm – Gastric + Brain	D43 (Neoplasm of uncertain or unknown behavior of brain and central nervous system) (0.81), D33 (Benign neoplasm of brain and other parts of central nervous system) (0.80), D32 (Benign neoplasm of meninges) (0.78)
Diseases	Top three similar ICD-10

I25 (Chronic ischemic heart disease) + I10 (Essential hypertension)	I51 (Complications and ill-defined descriptions of heart disease) (0.96), I27 (Other pulmonary heart diseases) (0.95), O13 (Gestational [pregnancy-induced] hypertension without significant proteinuria) (0.94)
H34 (Retinal vascular occlusions) + E11 (Non-insulin-dependent diabetes mellitus)	I67 (Other cerebrovascular diseases) (0.92), E13 (Other specified diabetes mellitus) (0.92), E12 (Malnutrition-related diabetes mellitus) (0.91)
COVID-19	A99 (Unspecified viral hemorrhagic fever) (0.94), J00 (Acute nasopharyngitis [common cold]) (0.92), B59 (Pneumocystosis) (0.92)

R00-R99 of ICD-10 indicates unclassified symptoms, signs, and abnormal clinical findings. We found the most similar ICD-10 code using an additive operation of up to three randomly selected R codes, using cosine similarity. This random selection was repeated 1,000 times. **Table 2** lists the top 10 results according to the cosine similarity. For example, a combination of R05 (cough) and R06 (abnormalities of breathing) is indicative of J459 (asthma, unspecified) with a cosine similarity of 0.971. A single symptom, R59 (enlarged lymph nodes), is found to be a symptom of C77 (secondary and unspecified malignant neoplasm of lymph nodes).

Table 2 | Examples of analogical reasoning with combinations of R codes. We randomly chose a combination of R codes and found the most similar ICD-10 code to the combination. The top 10 examples with the highest cosine similarity between the randomly chosen R codes and selected ICD-10 code are shown. We provided the code for constructing the full table (see Code availability).

R code 1	R code 2	R code 3	Most similar ICD-10
R06 (Abnormalities of breathing)	R11 (Nausea and vomiting)	R60 (Oedema, not elsewhere classified (NEC))	J81 (Pulmonary oedema)
R11 (Nausea and vomiting)	R29 (Other symptoms and signs involving the nervous and musculoskeletal systems)	R60 (Oedema, NEC)	M353 (Polymyalgia rheumatica)
R51 (Headache)	R73 (Elevated blood glucose level)		E162 (Hypoglycaemia, unspecified)
R19 (Other symptoms and signs involving the digestive system and abdomen)	R41 (Other symptoms and signs involving cognitive functions and awareness)	R59 (Enlarged lymph nodes)	D869 (Sarcoidosis, unspecified)
R05 (Cough)			J459 (Asthma, unspecified)
R29 (Other symptoms and signs involving the nervous and musculoskeletal systems)	R51 (Headache)	R60 (Oedema, NEC)	M353 (Polymyalgia rheumatica)
R12 (Heartburn)	R13 (Dysphagia)		K30 (Functional dyspepsia)
R05 (Cough)	R59 (Enlarged lymph nodes)	R64 (Cachexia)	D869 (Sarcoidosis, unspecified)
R59 (Enlarged lymph nodes)			C77 (Secondary and unspecified malignant neoplasm of lymph nodes)

R05 (Cough)	R06 (Abnormalities of breathing)	J459 (Asthma, unspecified)
-------------	----------------------------------	----------------------------

Comparing ICD2Vec pairs with disease–disease relationships

ICD2Vec was constructed based on the clinical words, sentences, and descriptions related to each disease. We sought to confirm the rationale and validity of ICD2Vec by comparing the biological relationships between diseases. We identified the significant associations between the cosine similarities obtained in our disease-to-disease pairs and three biological indicators, i.e., the relative risk for comorbidities in observational clinical data, separation in a disease-disease network, and genetic correlations based on genome-wide association studies.

Relative risk of comorbidity. The relative risk was quantified as the ratio of co-occurrence of the two diseases to the incidence of each disease. We only considered statistically significant relative risks. A total of 127,579 selected disease pairs are presented in **Fig. 2a**. The correlation of the top 10% of the cosine similarity group is higher than that of the total set of pairs (0.296 vs. 0.117). There is a tendency for a higher relative risk between two diseases with a higher cosine similarity. The mean difference in the relative risk between disease pairs with the top 10% of cosine similarity and the other group is statistically significant ($P = 2.6 \times 10^{-118}$, **Fig. 2d**).

Separation in a disease interactome network. A previous study by Menche *et al.* developed a network-based separation of two diseases as a measure of how close the diseases were in a disease–disease interactome network¹⁸. In other words, the smaller the value, the more similar the diseases. We considered a total of 33,332 pairs with separation values under a q-value of 0.05. The correlation of the top 10% of the cosine similarity group is negatively higher than that of the total set of pairs (-0.240 vs. -0.142) (Fig. 2b). The mean difference of the top 10% group is statistically lower than that of the other groups ($P = 7.6 \times 10^{-146}$, **Fig. 2e**).

Genetic correlation based on genome-wide association studies. A genetic correlation is a biological indicator that can be used to quantify disease-disease relationships. We compared the ICD2Vec cosine similarities with the genetic correlations of 1,531 pairs, as obtained using linkage disequilibrium score regression (LDSC) based on statistics of diseases from a genome-wide association study (GWAS). The genetic correlations between the top 10% cosine similarity group and the remaining groups are statistically different; the genetic correlations in the top 10% cosine similarity group are significantly higher than those in the other group ($P = 3.0 \times 10^{-5}$, **Fig. 2f**). In addition, there is a positive correlation between the genetic correlation and cosine similarity, although it is not statistically significant (**Fig. 2c**).

Fig. 2 | Comparison between cosine similarities of ICD2Vec pairs and biological relationships between disease pairs. We defined three factors to measure biological relationships between disease pairs: relative risk (a, d), separation (b, e), and genetic correlation (c, f). The factors were compared with the cosine similarity between ICD2Vec pairs. The red dots and bars represent the average of the pairs with the top 10% of cosine similarity, and blue dots and bars represent that of the other 90% of the pairs. We measured the correlations between the cosine similarities from the total and top 10% of ICD2Vec pairs and the biological relationships from corresponding disease pairs. The error bar indicates the standard error of the mean.

Use-cases: epidemiological studies using ICD2Vec risk score

We examined the association between a ICD2Vec risk score (IRIS) and the risks for four diseases: coronary artery disease (CAD), type 2 diabetes (T2D), dementia, and liver cancer. We designed prospective and retrospective cohort studies for each disease using data collected in the UK Biobank and with data extracted from the Clinical Data Warehouse (CDW) of the Samsung Medical Center (SMC) in South Korea. A description of these analyses is illustrated in **Supplementary Fig. 1**. The incidence of CAD, T2D, dementia, and liver cancer was defined using ICD-10 and other treatment codes (**Supplementary Table 1**). We collected the ICD-10 codes of patients who had been diagnosed with any disease at least once for a fixed duration before the start of the follow-up as the baseline. The detailed inclusion and exclusion criteria for selecting the patients in the two cohorts are described in **Supplementary Information**. The patients were grouped into deciles or 50-tiles according to the IRIS calculated for each patient.

We estimated the hazard ratios (HRs) for the IRIS values using a Cox proportional hazard model. The model was adjusted for age and sex. Additionally, we estimated HRs for baselines of 3, 5, 7, and 9 years in the UK Biobank and for a baseline of 3 years in SMC. The statistical significance was set at $P < 0.05$.

Table 3 shows the demographic characteristics and highest HRs for each target disease with the two independent datasets. For example, in the analysis for CAD using 9 years for the baseline and the top-three average of the IRIS values in the UK Biobank, 2,848 patients are diagnosed with CAD in the top 10% of the IRIS_{CAD} values, whereas 1,170 patients are diagnosed in the bottom 10% of the IRIS_{CAD} values. The effect of

the IRIS_{CAD} is significantly associated with the risk of CAD (adjusted HR, 2.08, 95% confidence interval [CI], 1.94-2.23). The highest HR for each disease was selected among eight HRs obtained from the two methods (R_{avg3} and R_{max}) for four baseline years (**Supplementary Tables 2–5**). Patients with a higher risk score tend to have a higher hazard for CAD (**Supplementary Fig. 2**). In the SMC, with 3 years as the baseline and the top-three average of the IRIS_{CAD} values, 1,532 patients are diagnosed with CAD in the top 10% of the IRIS_{CAD} values, whereas 437 patients are diagnosed in the bottom 10% of IRIS_{CAD} values (**Table 3**). The risk of CAD increases in SMC patients with higher IRIS_{CAD} values (adjusted HR, 2.41, 95% CI, 2.16–2.69). Similarly, the HRs for other target diseases also increase in individuals with higher IRIS_{CAD} values (**Supplementary Table 6** and **Supplementary Fig. 3**). Moreover, a higher incidence of CAD is observed in groups with high scores binned according to the 50-tiles of the IRIS_{CAD} (**Fig. 3**). For instance, the incidence of CAD in the highest risk score group is approximately four times higher than that in the lowest risk score group (19.2% vs. 5.0%).

Table 3 | Highest hazard ratios of ICD2Vec risk score for four diseases

Dataset	Target disease	Baseline (risk score derivation)	Top 10% of IRIS, [no. of samples, (no. of events)]	Bottom 10% of IRIS [no. of samples, (no. of events)]	Hazard ratio** (95% CI)
UK Biobank	CAD	9 years (top three average)	19,670 (2,848)	19,671 (1,170)	2.08 (1.94-2.23)
	T2D	9 years (top three average)	18,379 (2,230)	18,380 (971)	1.95 (1.81-2.11)
	Dementia	9 years (maximum)	19,791 (502)	19,792 (94)	3.49 (2.80-4.36)
	Liver cancer	7 years (maximum)	17,801 (588)	17,802 (162)	3.34 (2.81-3.98)
SMC	CAD	3 years (top three average)	13,664 (1,532)	13,665 (437)	2.41 (2.16-2.69)
	T2D	3 years (maximum)	13,036 (1,413)	13,037 (474)	2.62 (2.36-2.91)
	Dementia	3 years (top three average)	14,737 (1,030)	14,738 (109)	3.15 (2.57-3.85)
	Liver cancer	3 years (maximum)	14,790 (724)	14,790 (131)	5.98 (4.96-7.21)

* Hazard ratio was estimated using a Cox proportional hazard model adjusted for sex and age.

** All p-values for hazard ratios were estimated under 2×10^{-16} .

Abbreviations: Samsung Medical Center (SMC); coronary artery disease (CAD); type 2 diabetes (T2D); confidence interval (CI); ICD2Vec risk score (IRIS).

Fig. 3 | Incidence of four diseases according to IRIS risk groups. a, coronary artery disease (CAD). **b,** type 2 diabetes (T2D). **c,** dementia. **d,** liver cancer.

IRIS_{CAD} for a common disease and comorbid conditions

The IRIS, calculated based on each individual's disease history, not only identifies the risk groups for a single target disease but also provides insights into comorbid conditions.

To identify the comorbid conditions in a risk group based on the disease-wide risk score, we divided the study population at a baseline of 9 years into decile risk groups, as derived from the top three averages based on the onset of CAD. We used the 10 most common comorbidities for CAD and three other cardiovascular-related diseases as reported by the United States Department of Health & Human Services for comparison with actual prevalence²⁴. The 13 comorbidities for CAD generally show a higher prevalence among the high-risk CAD groups (**Fig. 4a**). Of the 13 comorbidities, the prevalence of hypertension has the strongest association with the high-CAD-risk group— 0.20% and 50.41% for the bottom 10% and top 10% CAD risk groups, respectively. In contrast, dementia has the lowest prevalence irrespective of the CAD risk group, with values of 0.01% and 0.37% for the bottom 10% and top 10% CAD risk groups, respectively. In addition, the prevalence of other comorbidities similarly increases for the top 10% CAD risk group relative to the bottom 10% group, including hyperlipidaemia (0.13–12.25%), T2D (0.66–8.39%), anaemia (0.03–8.36%), atrial fibrillation (0.10–7.25%), chronic obstructive pulmonary disease (0.01–5.10%), cataracts (2.03–4.15%), depression (0.36–3.70%), stroke (< 0.01–3.19%), heart failure (< 0.01–2.34%), rheumatoid arthritis (0.09–2.05%), and chronic kidney disease (< 0.01–1.94%) (**Fig. 4b**). The 20 most prevalent diseases in the highest CAD risk group are described in **Supplementary Table 7**. Either or without events, participants are most frequently diagnosed with hypertension (ICD-10: I10).

Fig. 4 | Prevalence of 13 comorbidities for CAD according to IRIS_{CAD} risk groups.

Abbreviations: type 2 diabetes (T2D); atrial fibrillation (AF); chronic obstructive pulmonary disease (COPD); heart failure (HF); rheumatoid arthritis (RA); chronic kidney disease (CKD).

We further investigated the risk of CAD among subgroups as divided by the IRIS_{CAD} and 10-year atherosclerotic cardiovascular disease (ASCVD) risk. The 10-year risk for ASCVD was categorized according to the definition by the American College of Cardiology Foundation²⁵. A high IRIS_{CAD} was associated with CAD risk, even within the same 10-year ASCVD risk group (**Table 4**). The cumulative CAD incidence of all other groups is higher than that of the 1st decile of IRIS_{CAD} values with a low 10-year ASCVD risk (**Supplementary Fig. 4**). In particular, the top 10% of IRIS_{CAD} values and high 10-year ASCVD risk groups show substantially increased CAD risks (adjusted HR, 4.65; 95% CI, 3.84–5.64).

Discussion

ICD-10 is a global standardized hierarchical code for representing diseases. In addition to its medical purposes, such as in clinical diagnosis, ICD-10 is essential in biomedical research because it is used to define disease variables⁸. In conventional analyses, such qualitative data are represented as dummy variables or one-hot encoding variables randomly distributed in the latent space; however, this representation loses the intrinsic relationships with other related variables. Accordingly, we proposed ICD codes that were vectorized to contain meaningful measurable quantitative values and suggested a

mathematical representation of diseases using ICD-10, called ICD2Vec. We showed that ICD2Vec contained the semantic and arithmetic relationships among diseases and presented medically conceptual meanings (**Tables 1, 2**). Moreover, the similarity of the vectors was significantly correlated with known biological relationships (**Fig. 2**). Furthermore, we found a significant association between the individual risk score derived from ICD2Vec and the actual disease incidence based on prospective cohort analyses of CAD, T2D, dementia, and liver cancer across two independent datasets from the UK and South Korea (**Supplementary Fig. 1** and **Table 3**), as discussed above. The higher risk score group for CAD (high IRIS_{CAD} carriers) had more CAD-related comorbid conditions (**Fig. 4** and **Table 4**).

Previous encoding techniques have used converted vectors to provide information^{19,26,27}. These techniques have also been applied to a variety of concepts in medicine and biology^{17,28}. In some studies, mathematical vectors embedding medical histories were used to predict the future statuses of patients^{29,30}. Applications to unstructured medical records provided novel insights for quantifying the relationships between diseases in multi-dimensional vector spaces²⁹. However, these algorithms have limitations; they need to be retrained for a specific dataset, and the resulting vectors are not consistent. In other research, vectors embedded with medical knowledge were used for a question-answering task^{16,31,32}. These algorithms showed improved clinical natural language processing (NLP) tasks. However, they were not used to convert ICD codes to vectors, although it is reasonable and useful to use vectorized ICD codes to analyse medical conditions. To the best of our knowledge, only one study has been conducted to elucidate the relationships between embedded vectors and clinical definitions³². The study compared their own vectors with medical concepts rated by physicians. Their vectors were validated as embedded vectors for readmission prediction based on a large cohort set. We quantitatively demonstrated that the vectors of ICD2Vec were meaningfully distributed in the latent space through a statistical investigation of the relationship between disease pairs based on three biologically established indicators (comorbidity as a relative risk, separation, and genetic correlation). We also confirmed the utility of ICD2Vec across the association study using the UK Biobank and the SMC cohort.

In addition, we defined individual risk scores based on ICD2Vec and performed use-case studies to investigate whether the scores for past diseases were effective in classifying risk groups. As a result, the risk of a target disease was predicted using these meaningful vectors. We observed a significant association between the risk of a target disease and the IRIS values for all other diseases during a specific baseline period (**Fig. 3**). In our experiments, the disease incidence of patients in the high-risk group was higher than that in the low-risk group (**Table 3**, **Supplementary Tables 1–4**, **Fig. 3**, and **Supplementary Figs. 2–3**). Specifically, we identified ICD codes related to the 10 most common comorbidities for CAD, which were previously reported as diseases with a significant association with CAD²⁴ (**Fig. 5**). Patients in the high-risk group were more likely to have underlying diseases relevant to the target disease over several years before the manifestation of the target disease. We confirmed that patients with a higher 10-year ASCVD risk and higher IRIS_{CAD} risk group were likely to develop CAD (**Table 4**). Notably, the IRIS index derived from ICD2Vec is intuitive and simple to calculate.

ICD2Vec is publicly available and applicable to various fields. Several examples of possible applications have been suggested. First, the IRIS values of all other diseases corresponding to a target disease can be computed and used as a risk predictor or covariate. Typically, researchers define arbitrarily selected relevant diseases as covariates or define an index for scoring whether participants were diagnosed with other diseases using the Charlson Comorbidity Index. Although conventional approaches adjust the main effect restrictively with only diseases selected and known by researchers, the IRIS is computed from all diseases and can be used to analyse a disease with unknown covariates. Second, a multimodal analysis requires efficient vectors for the features of a disease. The size of ICD2Vec can be lower than those of one-hot encoded ICD-10 codes. The dimensions (300) of ICD2Vec are reduced more than 60 times relative to the original dimensions (19,514) of the ICD codes. Third, using pretrained vectors, it is possible to increase the performance of a predictive model. Owing to the pretrained knowledge embedded in ICD2Vec, the parameters of a model can be well optimized by ICD2Vec, i.e., better than the approach with randomly defined variables used by one-hot encoding. Lastly, ICD2Vec might be helpful in mapping ICD standards to other standard terminologies, such as “Systematized Nomenclature of Medicine - Clinical Terms” and “Medical Subject Headings.” By calculating the cosine similarity between codes of different standards with an

algorithm such as ICD2Vec, code mapping can be performed in the order of the highest similarity and relevance.

Our study had several limitations. First, analogical reasoning with ICD2Vec does not represent a clinical diagnosis. Clinical diagnosis involves systematic clinical reasoning with patient information such as disease history, physical examination, and laboratory results³³⁻³⁵. For example, the key diagnostic factor of pulmonary oedema is excessive fluid, as revealed from medical lung imaging tests³⁶. Similar ICD codes are not the results of a diagnosis but have similar sources. Thus, in this study, we presented the top three similar ICD codes to illustrate the arithmetic and semantic relationships among the ICD2Vec codes. Second, the current ICD2Vec was based on FastText, which was trained on a general corpus rather than a medical corpus. Certain words need to be redefined to be more appropriate for medical use. For example, the “down” in Down Syndrome does not mean “being or moving lower” and should be used in Down Syndrome itself. Future work is required to update the vectors using a pretrained NLP model with a medical corpus. Third, converting ICD codes to vectors largely depends on the descriptions of the ICD codes and medical knowledge. Therefore, ICD2Vec needs to be updated as medical knowledge accumulates. Fourth, advanced algorithms such as BioBERT or Generative Pre-trained Transformer-3 can be used to generate vectors^{16,37}. A comparison of different algorithms is required to improve the performance of ICD2Vec.

ICD2Vec is a framework for converting qualitatively measured ICD codes into quantitative vectors containing the semantic relationships among diseases. The publicly available ICD2Vec can be used in diverse research and medical practices.

Methods

Dataset

We extracted the disease definitions from ICD-10-CM from a publicly available web database: ICD-10-CM is a clinical modification of ICD-10 in the United States and additionally extends some medical conditions or illnesses³⁸. The disease names from the ICD codes were used to encode the information of each code. Additionally, the clinical information and approximate synonyms of the ICD codes were incorporated into their vector representations. The clinical information was obtained from official medical websites, such as the National Institutes of Health of the United States and a summary in MedlinePlus provided by the Centers for Disease Control and Prevention. Incorporating this information into the vector representation enabled us to obtain a well-defined and accurate representation of the disease codes (**Fig. 1**). We used 321,593 words from 19,154 ICD codes and 1,166,960 words with clinical information and approximate synonyms.

Our embedded vectors were evaluated for generalization using a large prospective cohort from the UK Biobank and massive EHR data from the CDW of SMC. The UK Biobank provides approximately 500,000 participants' medical histories from interviews and questionnaires, as well as biological samples for genotyping and informed consent for long-term medical follow-up through links to national health registries. This research was conducted using the UK Biobank Resource under Application Number 33002. Second, the EHRs of over 200,000 patients who visited the SMC between 2010 and 2012 were collected in a hospital information system, i.e., the CDW of SMC. We could access and retrieve deidentified records, including demographic information and the diagnosis histories of the participants. Ethical approval for the study protocol was granted by the Institutional Review Board (No: 2021-07-016) at SMC.

Pre-processing for ICD2Vec

NLP is a subfield of computer science and a branch of artificial intelligence that helps computers to read, understand, interpret, and manipulate human languages. The word embeddings in NLP are a type of word representation in a high-dimensional space, such that words with similar meanings have similar representations and are positioned close to each other in the vector space. Word embeddings help the machine capture the context of human language. They are used to efficiently solve NLP problems owing to their ability to capture the semantic relationships among words.

We performed pre-processing using the following procedures: tokenization, stop-word removal, and stemming. Tokenization is the most common pre-processing task in NLP. Tokenization basically involves splitting up a corpus of text and breaking it up into individual words called tokens. Stop-word denotes a commonly used word, for example, “the,” “a,” “an,” or “in,” which does not add much meaning to the sentence and can be ignored without affecting the meaning of the sentence. It can be removed by storing a list of words considered to be stop-words. Here, this was accomplished using the Natural Language Toolkit (NLTK), a library in Python that includes a list of defined stop-words for the English language. Stemming is a text normalization technique process in NLP where words are reduced to their word stems. A stemming algorithm reduces the word “likes,” “liked,” “likely,” or “liking” to its base word, i.e., the word “like.” It is desirable for removing unnecessary noise, as often the base words and their inflected words mean the same thing. We performed stemming using the “Snowball Stemmer” from the NLTK.

After pre-processing the data, the clinical information and approximate synonyms from the ICD-10 website were combined with the ICD code information. This provided more defined descriptions of the disease codes and helped us to obtain more accurate vectors corresponding to each ICD code. In general, ICD codes have a parent–child hierarchical structure from the top level (e.g., A00) to the bottom level (e.g., A001). We added the clinical information and approximate synonyms of the ICD codes of the bottom level to the ICD codes of the top level, such as A00-Z99, to maintain the hierarchical structures.

ICD2Vec development

We used a pre-trained FastText model trained over a Wikipedia corpus to obtain vectors for the words in our prepared data. This pre-trained FastText model contained vectors with 300 dimensions obtained using the skip-gram model with the default parameters. The skip-gram model can be considered as a model for predicting the context, given a word. The input is the current word, whereas the output is the preceding words and the words following the current word. If the current word in the sentence is represented as w_i , the input to the model is w_i , whereas the outputs of the model are $w_{i-2}, w_{i-1}, w_{i+1}$, and w_{i+2} . After preparing the data, each word representing every ICD code was passed through the pre-trained FastText model. The FastText model provided a unique vector corresponding to each word.

A vector for each ICD code was obtained by averaging the vectors for the words in its description. For example, if the ICD code c_k contained n words $w_1, w_2, \dots, w_n$, we obtained vectors for each of these words denoted by $v_1, v_2, \dots, v_n$ after passing each of the words through the FastText model. Then, the vector for the ICD code was calculated by averaging the vectors for the words;

$$Vec(c_k) = \frac{\sum_i^n v_i}{n}.$$

IRIS

The most advantageous feature of ICD2Vec is that the disease definition, as a high-dimensional concept, can be easily expressed as an arithmetic vector. Because the vectors have meaningful relationships with each other, it is possible to adapt the vectors as a clinical variable for epidemiological or association studies. We suggested a disease-wide risk score for quantitatively measuring the risk of a target disease using ICD2Vec.

The disease-wide risk score (R_i) for a diagnosed disease calculated using the cosine similarity (CS) between the disease (D_i) and target disease (T) was defined as $R_i = CS(D_i, T)$. We defined the IRIS using the top three averages or using the maximum of the individual disease-wide risk score.

Definitions of biological factors

We considered the relative risk, network-based separation, and genetic correlation as quantitative biological measurements of the relationships between diseases.

Relative risk. The relative risk can be used as a measure of disease comorbidity. We analysed the co-occurrences of diseases within a patient. The relative risk RR is defined as follows¹⁸:

$$RR = \frac{n_{AB}n_{tot}}{n_A n_B}, \text{ where}$$

n_{tot} = the total number of patients in the data;

n_A, n_B = the number of patients diagnosed with disease A and B, respectively; and

n_{AB} = the number of patients diagnosed with both diseases A and B.

Additionally, the 99% confidence interval CI for the statistical significance is defined as follows:

$$CI = (RR \times \exp(-2.45\sigma), RR \times \exp(2.45\sigma)), \text{ with}$$

$$\sigma = \frac{1}{n_{AB}} + \frac{1}{n_A n_B} - \frac{1}{n_{tot}} - \frac{1}{n_{tot}^2}.$$

We calculated the risk of a disease relative to other diseases as the ratio of co-occurrence of the two diseases in the medical records of the UK Biobank.

Separation. A network-based separation was suggested that compared the network-based shortest distance between two diseases to the shortest distances within each disease¹⁸. This measurement indicates the degree to which two network-based disease modules are separated. A zero or negative value indicates that the two diseases may share pathologically similar clinical characteristics. The separation S_{AB} with disease A and B is defined as follows:

$$S_{AB} \equiv \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2}, \text{ where}$$

$\langle d_{AB} \rangle$ is the shortest distance between diseases A and B, and $\langle d_{AA} \rangle, \langle d_{BB} \rangle$ is the shortest distances within diseases A and B, respectively.

Genetic correlation. To examine the biological rationale of ICD2Vec, we compared the cosine similarities of ICD2Vec and genetic correlation values for each pair of diseases. The genetic correlation is an indicator that quantifies the genetic relationship between two traits, based on summary statistics from a GWAS. The genetic correlation ranges from -1 to +1, where -1 indicates a strong negative association and +1 indicates a strong positive association. Comparisons between cosine similarities of ICD2Vec and genetic correlations for pairs of diseases can provide useful etiological insights.

We used a total of 1,585 ICD-10-based UK Biobank GWAS summary statistics data from <https://www.nealelab.is/uk-biobank>. For each disease pair, we performed an LDSC on the pair of corresponding GWAS summary statistics³⁹. We considered disease pairs with positive correlation values and significance ($p < 0.05$).

Statistical analysis

We performed a Student's t-test and Pearson's correlation test to compare the correlations between the three biological factors and cosine similarity. We also measured the hazard risk of the target disease with individual risk scores using Cox proportional hazard regression. All statistical analyses were performed using R statistical software (version 3.6.1; R Foundation for Statistical Computing, Vienna, Austria). Statistical significance was declared at $P < 0.05$.

Data availability

Our ICD2Vec embedding dataset is available at <https://github.com/YeongChanLee/ICD2Vec>. We crawled the information on ICD-10 in <https://www.icd10data.com>. The UK Biobank GWAS summary statistics are available at <http://www.nealelab.is/uk-biobank>. The genotype and phenotype data can be obtained from the UK Biobank (<https://www.ukbiobank.ac.uk>) upon project application. The EHR data that support the findings of this study are available from the CDW of SMC but restrictions apply to the availability of these data, which were used under approval for the current study, and so are not publicly available.

Code availability

Code to construct ICD2Vec is available at <https://github.com/YeongChanLee/ICD2Vec>.

References

1. Parasrampur, S. & Henry, J. Hospitals' use of electronic health records data, 2015–2017. *ONC Data Brief* (2019).
2. Adler-Milstein, J. et al. Electronic health record adoption in US hospitals: the emergence of a digital “advanced use” divide. *Journal of the American Medical Informatics Association* **24**, 1142–1148 (2017).
3. Jetté, N. et al. The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. *Medical care*, 1105–1110 (2010).
4. Hirsch, J. et al. ICD-10: history and context. *American Journal of Neuroradiology* **37**, 596–599 (2016).
5. World Health Organization. International statistical classification of diseases and related health problems: Tabular list. Vol. 1 (World Health Organization, 2004).
6. Topaz, M., Shafran-Topaz, L. & Bowles, K. H. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. *Perspectives in health information management/AHIMA, American Health Information Management Association* **10** (2013).
7. Stausberg, J., Lehmann, N., Kaczmarek, D. & Stein, M. Reliability of diagnoses coding with ICD-10. *International journal of medical informatics* **77**, 50–57 (2008).
8. Zethelius, B. et al. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *New England Journal of Medicine* **358**, 2107–2116 (2008).
9. Gilbert, T. et al. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study. *The Lancet* **391**, 1775–1782 (2018).
10. James, M. T. et al. Derivation and external validation of prediction models for advanced chronic kidney disease following acute kidney injury. *Jama* **318**, 1787–1797 (2017).
11. Nielsen, A. B. et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *The Lancet Digital Health* **1**, e78–e89 (2019).
12. Park, J. H. et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ digital medicine* **3**, 1–7 (2020).
13. Sperandei, S. Understanding logistic regression analysis. *Biochemia medica* **24**, 12–18 (2014).
14. Suits, D. B. Use of dummy variables in regression equations. *Journal of the American Statistical Association* **52**, 548–551 (1957).
15. Howard, A. G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).
16. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
17. Solares, J. R. A. et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of biomedical informatics* **101**, 103337 (2020).
18. Menche, J. et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347** (2015).
19. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013).
21. Su, Y., Shen, Q. & Fan, X. Loop myopexy surgery for strabismus associated with high myopia. *Journal of ophthalmology* **2016** (2016).
22. Ekdawi, N. S., Nusz, K. J., Diehl, N. N. & Mohny, B. G. The development of myopia among children with intermittent exotropia. *American journal of ophthalmology* **149**, 503–507 (2010).
23. Mayo Clinic. Coronavirus disease 2019 (COVID-19). *Mayo Clinic* <https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963?p=1> (2020).

24. Arnett, D. K. et al. AHA/ACC/HHS strategies to enhance application of clinical practice guidelines in patients with cardiovascular disease and comorbid conditions: from the American Heart Association, American College of Cardiology, and US Department of Health and Human Services. *Circulation* **130**, 1662-1667 (2014).
25. Goff, D. C. et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology* **63**, 2935-2959 (2014).
26. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
27. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
28. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* **25**, 1419-1428 (2018).
29. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* **6**, 1-10 (2016).
30. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. in *Machine learning for healthcare conference*. 301-318 (PMLR).
31. Alsentzer, E. et al. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).
32. Huang, K., Altosaar, J. & Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).
33. Crombie, D. L. Diagnostic process. *The Journal of the College of General Practitioners* **6**, 579 (1963).
34. Diamond, G. A. & Forrester, J. S. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *New England Journal of Medicine* **300**, 1350-1358 (1979).
35. National Academies of Sciences, E. & Medicine. Improving diagnosis in health care. (National Academies Press, 2015).
36. Grainger, R. III. Interstitial Pulmonary Oedema and its Radiological Diagnosis. A Sign of Pulmonary Venous and Capillary Hypertension. *The British Institute of Radiology* **31**, 364 (1958).
37. Brown, T. B. et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
38. ICD10Data. The Web's Free 2021 ICD-10-CM/PCS Medical Coding Reference. *ICD10Data* <https://www.icd10data.com/> (2021).
39. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236 (2015).

Acknowledgements

This study was supported by a National Research Foundation of Korea Grant funded by the Ministry of Science and Information and Communication Technologies, South Korea (NRF-2019R1A2C4070496), and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), as funded by the Ministry of Health & Welfare, Republic of Korea (HI19C1132).

Author contributions

Y.C.L., S-H.J., and H-H.W. conceived and designed the study. Y.C.L., S-H.J., A.K., I.S., and M.S. collected the data. Y.C.L., S-H.J., A.K., and H-H.W. analysed and interpreted the data. Y.C.L., S-H.J., and A.K. did the statistical analysis. Y.C.L., S-H.J., and A.K. wrote the manuscript. I.S., M.S., M.S.K., K.K., W.M., and W-Y.P. read and provided critical revision of the manuscript for intellectual contents; and all authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ICD2VecSupplementaryInformation.docx](#)
- [figureS1.pdf](#)
- [figureS2.pdf](#)
- [figureS3.pdf](#)
- [figureS4.pdf](#)