Unified Clinical Vocabulary Embeddings for Advancing Precision Medicine

Ruth Johnson[1,2], Uri Gottlieb[3], Galit Shaham[3], Lihi Eisen[3], Jacob Waxman[3], Stav Devons-Sberro[3], Curtis R. Ginder[4], Peter Hong[5,6,7], Raheel Sayeed[2], Ben Y. Reis[1,2,8,11], Ran D. Balicer[1,3,9], Noa Dagan[1,2,3,10,*], and Marinka Zitnik[1,2,11,12,13,*]

[1] The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute, Boston, MA, USA
[2] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[3] Clalit Research Institute, Innovation Division, Clalit Health Services, Ramat-Gan, Israel
[4] Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[5] Division of General Pediatrics, Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA
[6] Information Technology, Enterprise Data Analytics and Reporting, Boston Children's Hospital, Boston, MA, USA
[7] Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[8] Predictive Medicine Group, Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
[9] Faculty of Health Sciences, School of Public Health, Ben Gurion University of the Negev, Be'er Sheva, Israel
[10] Software and Information Systems Engineering, Ben Gurion University, Be'er Sheva, Israel
[11] Harvard Data Science Initiative, Cambridge, MA, USA
[12] Broad Institute of MIT and Harvard, Cambridge, MA, USA
[13] Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA, USA
[*] Joint correspondence: noada@clalit.org.il and marinka@hms.harvard.edu

Integrating clinical knowledge into AI remains challenging despite numerous medical guidelines and vocabularies. Medical codes, central to healthcare systems, often reflect operational patterns shaped by geographic factors, national policies, insurance frameworks, and physician practices rather than the precise representation of clinical knowledge. This disconnect hampers AI in representing clinical relationships, raising concerns about bias, transparency, and generalizability. Here, we developed a resource of 67,124 clinical vocabulary embeddings derived from a clinical knowledge graph tailored to electronic health record vocabularies, spanning over 1.3 million edges. Using graph transformer neural networks, we generated clinical vocabulary embeddings that provide a new representation of clinical knowledge by unifying seven medical vocabularies. These embeddings were validated through a phenotype risk score analysis involving 4.57 million patients from Clalit Healthcare Services, effectively stratifying individuals based on survival outcomes. Inter-institutional panels of clinicians evaluated the embeddings for alignment with clinical knowledge across 90 diseases and 3,000 clinical codes, confirming their robustness and transferability. This resource addresses gaps in integrating clinical vocabularies into AI models and training datasets, paving the way for knowledge-grounded population and patient-level models.

Introduction

Medicine is built on centuries of knowledge and the pursuit of individualized patient care through meticulous reasoning and evidence-based practice[1]. Over time, numerous medical vocabularies and ontologies have been developed to represent clinical information, fostering interoperability and global data exchange[2–4]. However, while these standardized coding systems provide a consistent framework for representing clinical knowledge, they remain fragmented, with each vocabulary optimized for specific purposes. This lack of unification across vocabularies poses challenges for their integration into artificial intelligence (AI) models and training datasets, creating a gap in the effective use of clinical knowledge for precision medicine.

The increasing adoption of electronic health records (EHRs) and data standardization has driven medicine toward data-driven approaches[5,6], with more than half of healthcare foundation AI models now relying exclusively on structured clinical codes, such as billing data and medication records[7]. These models generate predictions by identifying statistical patterns and latent representations of patients and patient populations from EHR datasets. However, this process assumes that datasets contain the requisite information to represent clinical concepts and that inferred associations align with clinical knowledge. This assumption poses challenges when models capture patterns unique to specific healthcare settings rather than generalizable insights[8–11]. This issue is pronounced in prediction models trained on clinical codes from structured EHR data, where high variability across institutions can limit model generalizability[12–15]. Clinical codes are often optimized for operational needs[16], encoding patterns shaped by geographic factors[17–19], national policies[20–23], health insurance[24–26], physician practices[27–29], and other factors that affect healthcare delivery[30–34]. EHR-based models, as a result, often fail to capture generalizable definitions and dependencies among clinical codes, which are critical for effectively leveraging clinical knowledge[35–39].

The lack of alignment between EHR-based prediction models and clinical knowledge also introduces concerns about model bias and transparency[40]. Although large language models (LLMs) offer implicit access to clinical knowledge embedded within their training data, unvetted datasets from biological data repositories and scientific articles used to train LLMs can increase the risk of bias and mistakes in generated outputs[41–43]. Additionally, LLMs struggle to effectively represent clinical codes[44–46], especially rare or highly specific codes, resulting in inaccuracies, over-generalizations, or hallucinations during inference[47]. This can be concerning given that a large portion of medical codes are highly specific and may not be used in everyday clinical practice where, for example, fewer than 5% of SNOMED CT codes account for 95% of usage in healthcare institutions[48], creating a mismatch between what is modeled and what is clinically relevant.

There is a gap in AI in integrating clinical knowledge and leveraging clinical expertise to produce more reliable and accurate predictions[49,50]. Clinical researchers have distilled scientific evidence into precise recommendations, resulting in over 3,700 guidelines published across 39 countries[51–54]. Despite the critical importance of these guidelines, most AI models lack mechanisms to effectively and systematically integrate clinical knowledge. Instead, they rely on statistical associations derived exclusively from patient datasets or large text corpora, which struggle to capture the relationships encoded in structured EHR data. Bridging this gap requires consistent, machine-readable knowledge representations that map connections between clinical codes in EHRs and the broader clinical knowledge they reflect. While data types, such as text and images, can be transformed into high-dimensional embeddings to support AI models[55–60], there is no equivalent resource for representing clinical knowledge tied to structured medical data.

To address this need, we develop a resource that constructs embeddings for 67,124 medical codes, defining a unifying latent space of clinical knowledge. Using state-of-the-art relational graph transformers and a clinical knowledge graph, we create a cohesive, machine-readable map that captures relationships among seven clinical vocabularies, including laboratory tests, diagnosis codes, and medications, without requiring manual curation. By integrating verified knowledge bases and medical ontologies into a knowledge graph of standardized EHR codes, this resource reduces the risk of propagating inaccuracies while promoting transparency.

Our resource provides a hypothesis-free approach to generating clinically insightful representations of medical codes. It offers three main applications: (1) integrating clinical knowledge into precision medicine patient models, (2) enabling generalizable models of populations and patient subtypes that can be safely exchanged across institutions, and (3) providing insights into the organization of clinical knowledge. The latent space of medical codes reveals patterns consistent

with human anatomy and disease presentations, capturing symptomatic and clinical presentations of diseases that can be decomposed into symptom-level embeddings. We demonstrate the predictive utility of these embeddings through a large-scale phenotype risk score analysis for three chronic diseases across 4.57 million patients from Clalit Healthcare Services. An expert clinical evaluation across 90 diseases and 3,000 clinical codes conducted with clinician panels in the United States and Israel validates the alignment of these embeddings with established medical knowledge. Our findings establish unified medical code embeddings as a foundational resource for advancing AI-driven healthcare. Unified clinical vocabulary embeddings can facilitate collaborative, scalable efforts in clinical AI and deepen our understanding of disease mechanisms, laboratory tests, diagnosis codes, medications, and their underlying dependencies.

Results

Overview of Approach

We develop a map of clinical embeddings using a clinical knowledge graph (KG) specifically constructed for clinical vocabularies used in EHRs. We encode structural and relational information about each clinical concept from the KG in embeddings using state-of-the-art graph transformer neural networks. The construction of this resource can be broken down into three stages: (i) Integrating standard clinical vocabularies and existing databases into a unified knowledge graph (PheKG); (ii) Using a graph transformer neural network to model relationships between clinical codes; (iii) Learning joint representations of clinical codes through self-supervised learning and noise-contrastive estimation.

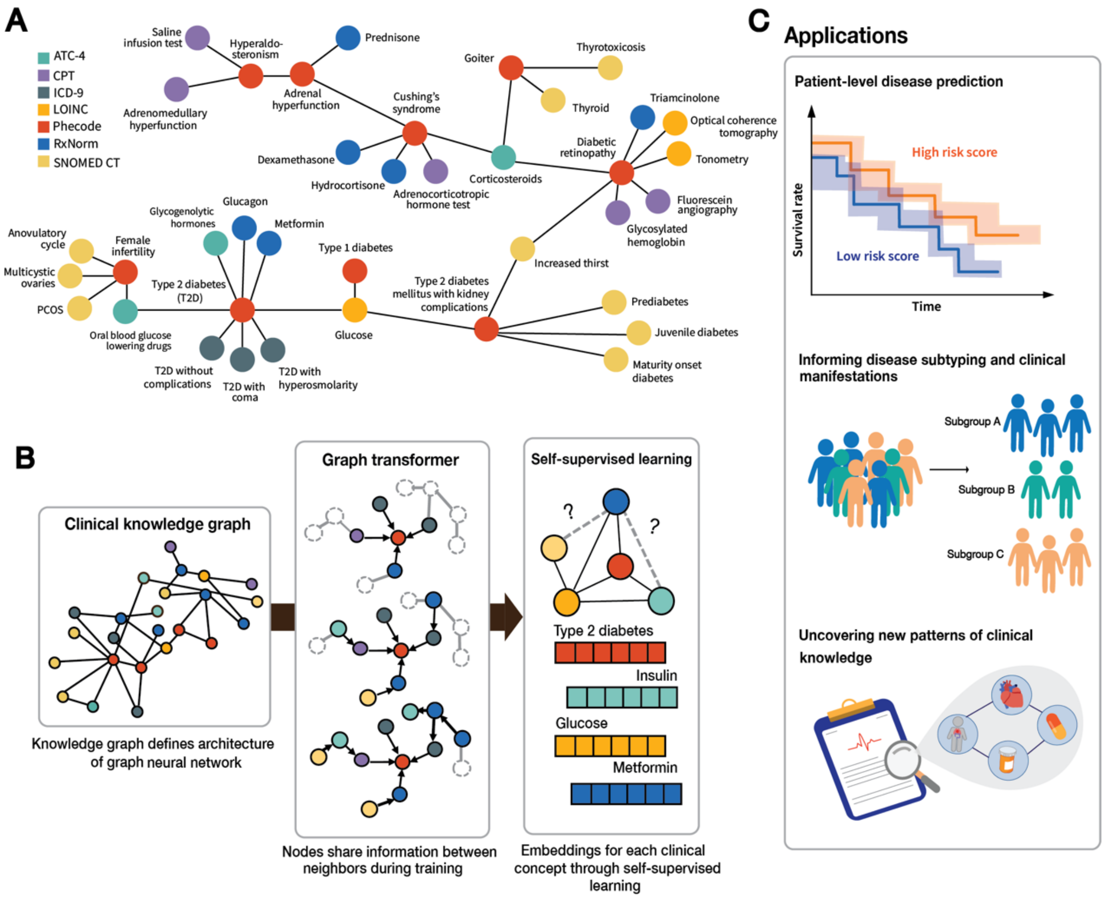


Figure 1: Overview of approach. (A) Illustrated is a local neighborhood of PheKG clinical knowledge graph. Each clinical concept is represented as a node, and the color denotes the node type (clinical vocabulary). **(B)** Workflow for generating clinical knowledge embeddings: building the clinical knowledge graph, constructing a graph transformer model, and generating embeddings through self-supervised learning. **(C)** Use cases of unified clinical vocabulary embeddings.

Knowledge graphs offer an intuitive approach for structuring biomedical knowledge due to their inherent capability to organize data across hierarchies and accurately capture intricate relationships. Although many biomedical KGs have been constructed, most do not explicitly incorporate standard clinical vocabularies. Instead, they focus on biological processes[61–63], such as genes and pathways, therapeutics[55,64,65], genetic phenotypes[66–69], or a combination thereof[70–72]. Furthermore, KGs that incorporate clinical vocabularies are often restricted to ICD (diagnosis codes) or rely on medical terminologies like HPO[73] and MONDO[74], which are not standard in most EHRs[75–77]. KGs can also be constructed directly from patient data[78–80]. However, this method is highly sensitive to the selection of the patient population and risks capturing health patterns specific to the chosen group, limiting the generalizability of the resulting knowledge graph.

To maximize the portability of PheKG, we construct the clinical knowledge graph using medical vocabulary without incorporating any patient information. We start from PheMap, a knowledge base of clinical concepts based on the natural language processing of biomedical literature[81]. For a given phecode (an aggregated version of diagnostic codes[82]), PheMap provides the pairwise importance score between a given phecode and clinical codes from standardized medical vocabularies, including ICD-9, SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms), RxNorm, and LOINC codes, yielding an initial set of 78,801 codes and 727,939 pairwise-relationships. Next, we incorporate drug information from the Anatomical Therapeutic Chemical Classification (ATC)[83]. The RxNorm[84] vocabulary is a terminology for medications and includes information regarding dosage and ingredients, and the ATC system is organized according to drug therapeutic and chemical properties. We use the Unified Medical Language System (UMLS) database to integrate this information into the knowledge base to identify the ATC codes corresponding to each RxNorm code (Methods). The resulting knowledge base spans seven standardized vocabularies, all EHR coding systems (Figure 1A).

To construct the knowledge graph, we designate each clinical code as a node and the node type as the source vocabulary. Using the precomputed importance scores from PheMap, we create an edge between a phecode and a clinical code whenever the pair is assigned a nonzero importance score. To incorporate the hierarchical structure of the ontologies, we add edges between clinical codes to reflect the parent-leaf node relationships. To reduce redundancies within the graph, we merge SNOMED CT codes with their shared synonyms and LOINC part codes that share the same part name (but have different part types) and then aggregate their edges (Methods). Some codes in the knowledge graph (e.g., “bacterial infections”) are inherently broad and non-specific, resulting in an exceptionally high number of connections in the KG. These high-degree nodes can introduce noise and hinder representation learning models. We applied selective pruning to nodes with a degree >1,000 to mitigate this issue. We remove edges connecting these nodes to leaf-level ICD-9 codes (codes with two decimal places) while retaining connections to their parent-level ICD-9 codes. This approach preserves the hierarchical structure of the graph while reducing redundancy and noise. The resulting knowledge graph, PheKG, comprises 67,124 nodes spanning seven node types and 1,315,610 edges (Supplementary Table S1 and Methods).

Using a heterogenous graph transformer[85], we learn a function that maps each node (clinical code) in the KG to its low-dimensional representation (embedding) (Figure 1B). The KG determines how information is shared between nodes throughout the KG, while a multi-head attention mechanism enhances the model's ability to simultaneously up-weight and down-weight different parts of the topological space[86]. The final model architecture utilizes four graph transformer layers comprised of over 34 million parameters. We implement a self-supervised approach by forming the training objective as a supervised task where the labels are derived from the KG[87]. We train the model for edge prediction through self-supervised learning via contrastive edge masking[88] (Methods). This approach generates versatile representations for each clinical code, ensuring that codes proximal in the knowledge graph are likewise closely aligned in the latent space, reflecting their semantic similarity.

The central output of our approach is a universal embedding space that captures relationships between clinical codes across diverse medical vocabularies, represented by 67,124 learned clinical knowledge embeddings. We evaluate how well the latent space reflects human physiology, patterns of disease presentation, and performance in patient prediction tasks (Figure 1C). By leveraging a heterogeneous knowledge graph model, this approach integrates various clinical elements, including laboratory tests, procedures, and medications. This integration enables comparisons and connections between codes from different medical vocabularies, analogous to how health records chronicle a patient's medical history through a diverse range of documented healthcare interactions.

Clinical embeddings capture knowledge of human anatomy and clinical subspecialties

For the clinical knowledge embeddings to be helpful, they must recapitulate clinically meaningful patterns consistent with human biology, disease presentations, and healthcare practices. A patient's medical record comprises many clinical vocabularies, with various types of codes describing their medical history. A key advantage of joint embedding space is its ability to directly compare clinical codes across different vocabularies, bridging the gaps between disparate coding systems. We first aimed to characterize the breadth and granularity of the clinical knowledge represented in the learned embeddings by visualizing the latent space (Figure 2A). Examining the clusters of phecode embeddings, we find that phenotypes broadly group by organ type, mirroring the organization of clinical care (adjusted rand index (*ARI*): 0.27, adjusted mutual information (*AMI*): 0.40, *silhouette-score*: 0.34, computed over ten disease classes). Although some phenotypes deviate from these main clusters, many deviations reflect known distinctions within clinical care. For example, most endocrine conditions cluster together (Fig 2A: orange cluster), but some phenotypes cluster more closely to sense organ conditions (Fig 2A: green cluster). However, this small set of endocrine phenotypes includes mainly eye-related conditions resulting from complications from endocrine disorders, such as diabetic retinopathy and exophthalmos.

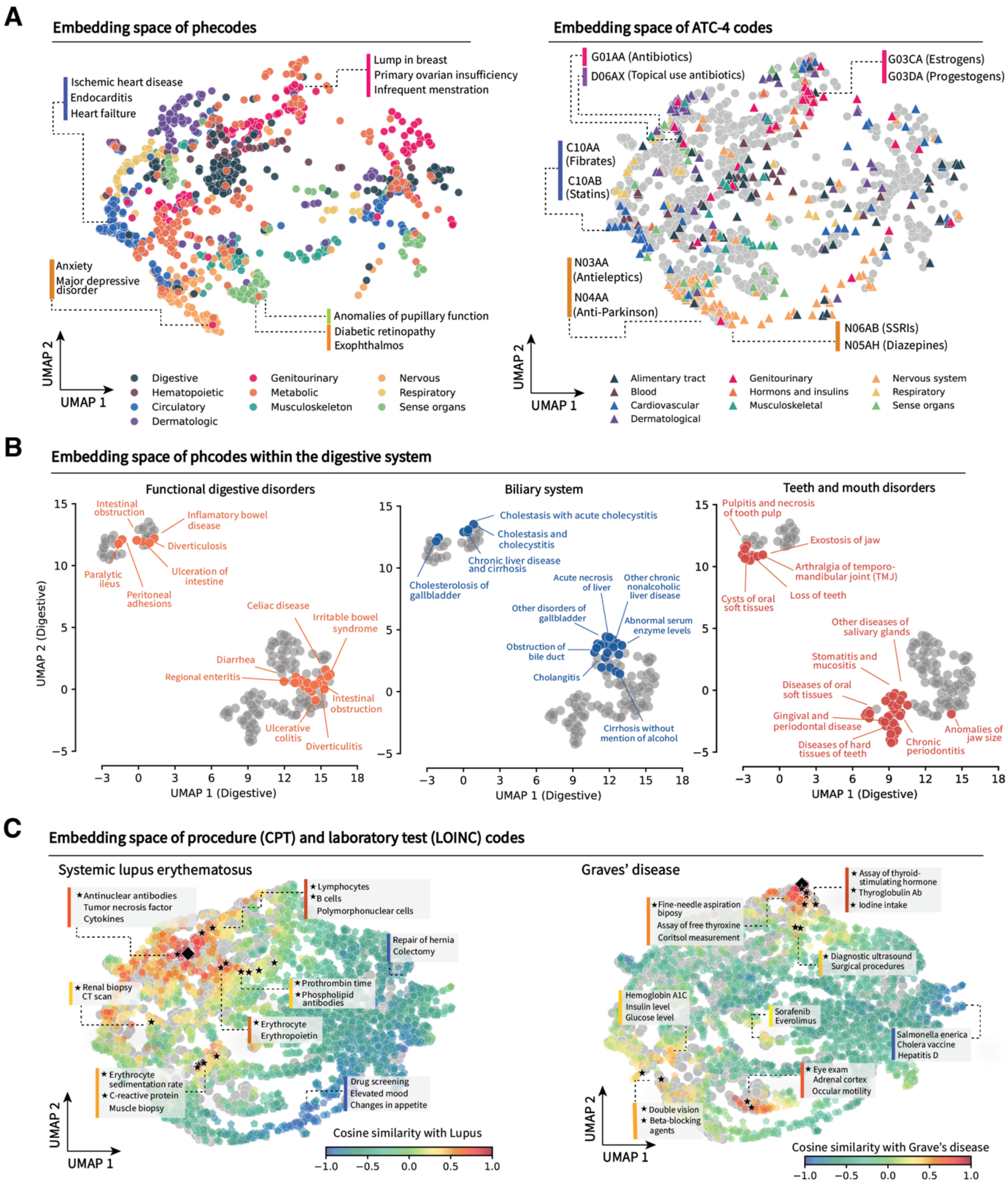


Figure 2. Latent space captures biomedical knowledge consistent with known organ system biology and anatomy. (A) Scatter plots showing a reduced dimension of the embedding space of clinical codes. On the left, the latent space of phecodes is shown, where each phecode is shaded by its corresponding organ system categorization. On the right, the latent space of ATC-4 medication codes is shown, where each ATC-4 code is shaded according to the corresponding ATC level 1 and encodes are shaded in gray. **(B)** Visualization of the embedding region of digestive system phecodes. We highlight the following phenotype subcategories: functional digestive disorders – phecodes

560-570 (left), biliary system – 570-580 (center), and teeth and mouth disorders - 520-530 (right). **(C)** Visualization of embedding space overlaid with LOINC part-codes and CPT codes. Codes are shaded based on cosine similarity with systematic lupus erythematosus (left) and Graves' disease (right). The target disease embedding is denoted with a black diamond. The black stars indicate a subset of the laboratory tests and procedures formally mentioned in each disease's diagnostic guidelines[89,90]. These are presented along with other codes in their proximity but are not mentioned directly in the guidelines.

Projecting the embeddings of ATC-4 codes into the same low-dimensional space of the diagnostic codes, we observe a striking parallel between the organization of phecodes and ATC-4 codes. Groups of ATC-4 codes loosely cluster according to ATC level 1 (ATC-1) classification, which organizes drugs according to anatomical and pharmacological group[83] (*ARI*: 0.16, *AMI*: 0.21, *silhouette-score*: 0.23, computed over ten drug classes). Broadly, codes clustered by ATC-1 are located near phenotypes from the corresponding organ system (Figure 2A). For example, the ATC-4 code N06AB (selective serotonin reuptake inhibitors; SSRIs) is situated near the phenotypes for anxiety disorder, major depressive disorder, and dysthymic disorder, which are consistent with the drug indications of SSRIs[91]. Again, we also find that deviations in the expected cluster labeling according to the ATC-1 classification reflect known artifacts of the ATC coding system. For example, the embeddings for ATC-4 codes G01AA (antibiotics) and D06AX (antibiotics for topical use) are located nearby in the embedding space despite being in different ATC-1 categories. However, given that both classes of drugs have overlapping disease indications, their proximity in the latent space is consistent with their known clinical usage.

To achieve maximum utility, the unified clinical vocabulary latent space must capture patterns at broad anatomical levels while also identifying discernible relationships at clinical subspecialty levels. To illustrate the granularity of clinical vocabulary embeddings, we perform a separate dimensionality reduction analysis using only phecodes of the digestive system (Figure 2B). We find that conditions within the same subspecialty are significantly closer together compared with all other conditions of the digestive system, where codes describing functional digestive disorders (*p-value*=3.65×10^{-45}; two-sided Mann-Whitney U test), biliary system (*p-value*=6.38×10^{-44}), and conditions of mouth and teeth (*p-value*=1.05×10^{-74}) are separated within the latent space. Although these codes all fall under the practice of gastroenterology, the embeddings exhibit distinct clustering patterns that reflect their unique clinical presentations and organ systems. This finding highlights the ability of the clinical knowledge embeddings to capture fine-grained distinctions within broader medical categories, offering insights that can enhance both clinical knowledge and provide new insights.

Latent embedding space reflects patterns of disease presentation and diagnostic processes

Although we have exhibited how the embedding space captures broader clinical patterns, most downstream translational tasks will also require granularity at the disease level. To assess how patterns of a given disease are encoded throughout the latent space, we use Graves' disease and systemic lupus erythematosus as two case studies (Figure 2C). For each disease, we compute the cosine similarity between the disease embedding and each clinical code within the embedding space and visualize these patterns by projecting the computed cosine similarity onto the latent space. Broadly, as embeddings move closer to the target disease embedding, the cosine similarity increases where we observe regions of high cosine similarity with lupus and Graves' disease embeddings, as shaded in red and orange.

Highlighting procedures and laboratory tests used in the diagnostic process of lupus[90], we find that many codes within this region are consistent with known lupus diagnostic guidelines, including tests for anti-nuclear antibodies and lymphocytes and B-cells (Methods). For Graves' disease, we observe a similar pattern where lab and procedure embeddings with a high cosine similarity score are located near the disease embedding within the latent space. Many of these align with the diagnostic guidelines for Graves' disease, including the codes for an assay of thyroid stimulating hormone, thyroglobulin antibody test, and fine needle aspiration biopsy[89]. This clustering reflects how the unified embedding space captures the relationships between different types of clinical codes, providing a cohesive framework for connecting diagnoses, procedures, and tests across clinical subspecialties.

Furthermore, we observe distinct clusters showing high cosine similarity between clinical codes and each disease despite being separate from the region surrounding the disease embedding. For lupus, this region includes tests such as erythrocyte sedimentation rate and C-reactive protein, both commonly used to assess systemic inflammation. Similarly, we identify a cluster of eye-related tests for Graves' disease, including codes for eye exams and ocular motility. This highlights the embedding space's ability to capture nuanced patterns that extend beyond identifying codes associated with the same organ system, recapitulating patterns aligned with established disease presentations.

Finally, to quantitatively assess the quality of the embedding space against known clinical knowledge of specific diseases, we compare the similarity between diseases and symptoms based on the guidelines described in the Phenotype Knowledgebase (PheKB), a database of clinically validated EHR-based phenotyping algorithms[92]. Although different algorithms within PheKB use various data types, such as clinical notes and structured EHR features, many utilize diagnostic codes. We select three diseases—chronic kidney disease, anxiety, and breast cancer—and measure the mean cosine similarity between the embedding of a given disease (in the form of a phecode embedding) and the embeddings of ICD codes corresponding to those listed in PheKB. We compare the mean similarity with a randomly selected set of ICD codes representing negative controls. Across the three diseases, we find that the group of codes derived from PheKB demonstrates significantly higher similarity than those from the control group (*mean p-value*: 6.15×10^{-5}; two-sided Mann-Whitney U test), meaning that clinical codes related to each phenotype are indeed located in proximity to one another in the embedding space (Supplementary Figure S1). The clinical knowledge embeddings align with established medical ontologies and reflect real-world diagnostic practices, offering versatility in examining different diseases.

Embedding arithmetic reveals the composition and divergence of disease manifestations

Clinical vocabularies have semantic properties that allow them to describe a wide range of clinical scenarios. Our research demonstrates that clinical knowledge embeddings can achieve comparable descriptive power through embedding arithmetic. This technique allows embeddings to be combined mathematically to form new meanings. For example, adding the embeddings for *diabetes* and *kidney disease* might result in an embedding closely representing *diabetic nephropathy*, effectively capturing the relationship between these conditions without explicitly defining it in the original vocabulary. Such compositions of disease symptoms, in the form of clinical code embeddings, produce embeddings that reside close to the target disease embedding within the latent space. In another example, the similarity between the vector summation of common heart disease symptoms (*e.g., shortness of breath, chest pain)* and the disease embedding of *common heart disease* yields a cosine similarity score of 0.66 (Supplementary Table S3). These observations mirror the geometric properties of word embeddings in natural language processing[93]. We refer to the aggregation of symptom embeddings in this manner as a *disease symptom embedding*—denoting that the resulting vector is a composition of the symptoms of a given disease (Figure 3A).

We assess this pattern by examining nine diseases with significant health implications across different clinical specialties. Five conditions are among the top non-communicable diseases with the highest mortality rate identified by the World Health Organization[94] (chronic obstructive pulmonary disease (COPD), Alzheimer's, heart disease, stroke, and lung cancer), and four are among the most common autoimmune disorders in adult populations (Graves' disease, multiple sclerosis, lupus erythematosus, and Crohn's disease)[95]. Based on the clinical descriptions for each disease described by the Mayo Clinic[96–100] and the manual translation of symptom lists to diagnosis codes, we aggregate all disease symptoms in the form of ICD-9 code embeddings (Supplementary Table S2).

Computing the cosine similarity between the disease symptom embedding and the target disease embedding across all nine diseases, we find that all pairs have positive similarity scores (*mean-cos-similarity*: 0.77, *SD*: 0.13) (Supplementary Table S3). Moreover, conditions with overlapping symptoms also exhibit high similarity scores. For instance, the phecode embedding for multiple sclerosis also has a high similarity score with the disease composition embedding for lupus (*cos-similarity*=0.58), as both diseases share symptoms such as fatigue and muscle pain. We observe high levels of congruence between diseases affecting the same organ system, such as lung cancer (target disease) and COPD (symptom embedding), which both affect the respiratory system (*cos-similarity*=0.86). We also observe divergent patterns when the symptoms and target disease differ. For example, computing the cosine similarity between each disease embedding and the symptom embedding for Crohn's disease, we observe low similarity scores across all the other conditions (*mean-cos-similarity*: -0.065, SD: 0.23). This may be because most symptoms of Crohn's disease affect the digestive tract, unlike most symptoms of the other eight diseases.

After observing this trend, we assess whether individual symptoms possess meaningful syntactic and semantic patterns within the embedding space. As a proxy for the disease specificity of each symptom, we order symptoms from highest to lowest frequency as measured within a sample of patients from Clalit Healthcare Services (CHS). Incrementally aggregating symptoms one by one for each of the nine diseases reveals striking patterns. As we add less common symptoms in the general patient population, the resulting pooled representation of symptoms gradually becomes more similar to the target disease embedding (Supplementary Figures S2, S3). We evaluate the robusticity of this effect by perturbing different elements of the embedding composition framework. Even after varying the target disease and symptom, sampling related symptoms, and adding randomly selected symptoms (as happens with patients), the symptom embedding still moves towards target disease embeddings in the latent space (Supplementary Figure S4). This analysis also illustrates that we can model meaningful patterns of diseases not explicitly provided in the knowledge graph, emphasizing the versatility of clinical knowledge embeddings.

This process can help study diseases with similar symptoms by revealing subtle differences in clinical presentations. This level of granularity can be observed when comparing various autoimmune diseases due to the presentation of overlapping symptoms. For Graves' disease, we compute the initial cosine similarity between the disease embedding for each of the four diseases and the first symptom embedding, "palpitations", (the Graves' symptom found to be most prevalent in the sampled patient population), but no disease has a cosine similarity score > 0.40 (Figure 3B). Given the relatively high prevalence of this symptom in the general patient population and lack of disease specificity (*freq*=0.232), it is unsurprising that initial symptoms do not show a high correlation with any of the immune conditions. However, as we add symptoms that are relatively less common within the general population, such as 'goiter' (*freq*=0.029), the composition of symptoms becomes increasingly similar to the representation of Graves' disease, and the similarity with the other three autoimmune conditions begins to diverge.

Analyzing the similarity scores of the disease embedding for Graves' disease with individual symptoms, most symptoms show lower similarity than when compared to the combined disease symptom embedding, and even three symptoms show negative similarity scores (skin/integumentary tissue symptoms, digestive symptoms, and fatigue) (Figure 3C). This analysis highlights the importance of considering the combination of symptoms instead of viewing each independently. These patterns mirror how individual symptoms may not provide sufficient information to assess the probability of a specific condition, but as a patient accumulates more symptoms over time, the aggregation of these symptoms provides a meaningful tool that can potentially stratify patients according to disease risk. Finally, as the embeddings used in this analysis are not based on patient-level data, the predictive insights derived from the aggregated symptom embeddings can be generalized and reliably used across clinical settings.

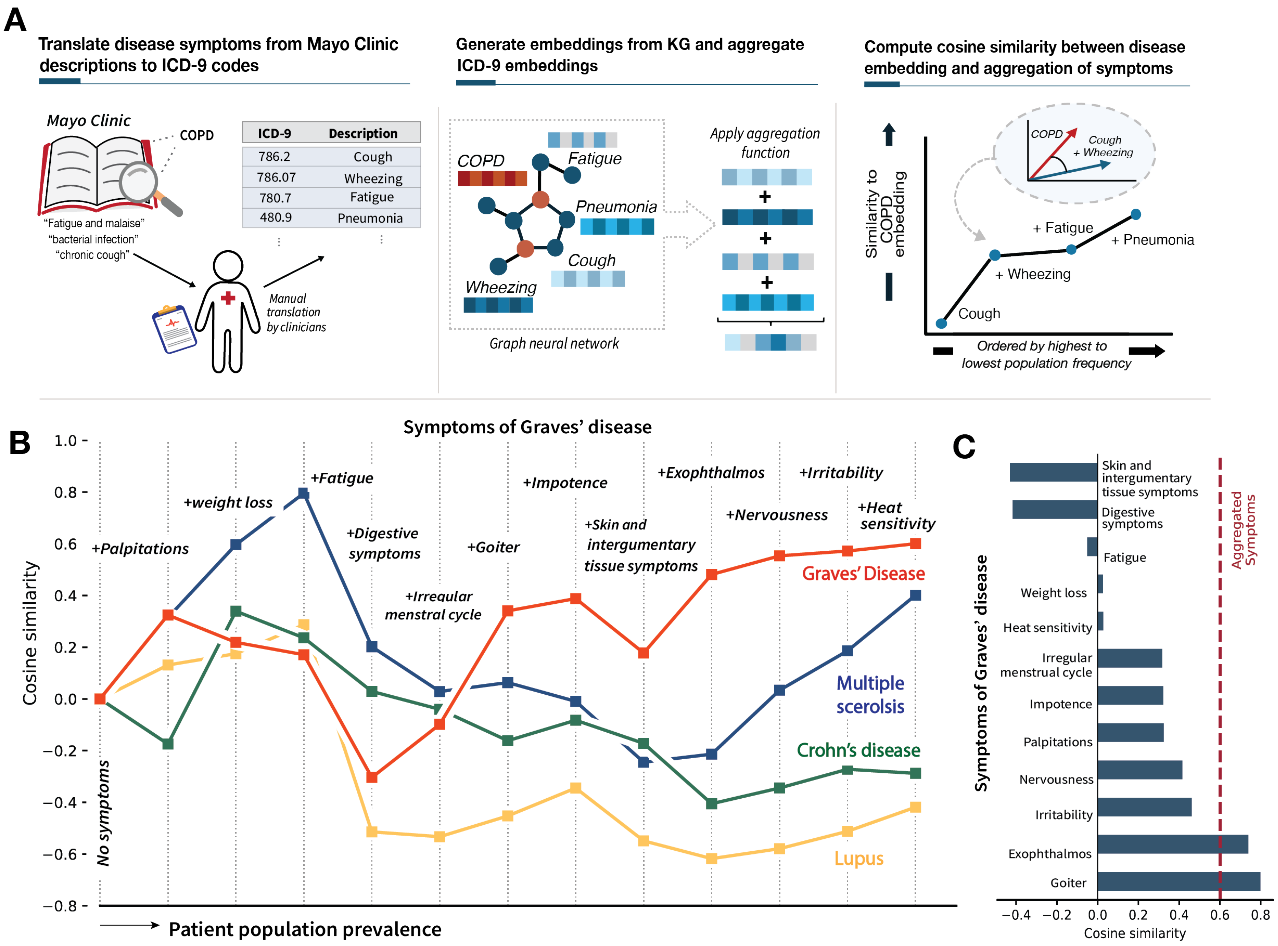


Figure 3: Embedding operations mirror symptomatic manifestations of clinical conditions. **(A)** Overview of approach for constructing disease symptom embeddings. **(B)** Line plot showing the cosine similarity across four autoimmune diseases and the aggregation of Grave's disease symptoms. Aggregated embeddings are computed by performing a vector sum across the embeddings for each symptom. **(C)** Bar plot showing the cosine similarity between Graves' disease and each symptom without aggregation.

Unified vocabulary embeddings enable disease risk stratification and severity prediction

Many clinical AI models focus on patient outcome prediction, and we explore the predictive utility of clinical knowledge embeddings in this context. Next, we investigate how these embeddings can be integrated into AI models for tasks such as predicting disease progression and disease risk. We compute a disease-specific risk score for various conditions and the relative disease risk for each patient. We only use information derived from the clinical knowledge embeddings to prevent the performance from being biased towards a specific patient-level training dataset. Analogous to polygenic risk score frameworks[101], we compute a phenotype risk score[102] by scanning a patient's EHR and aggregating the embeddings across conditions with the highest similarity to the disease of interest (Methods).

We validated our approach using retrospective CHS data from EHR data from 4.57 million individuals[103]. We utilized patients listed in the CHS chronic registry, a registry maintained by CHS that monitors individuals with specific chronic conditions[104]. We chose to focus on three clinical conditions with the largest sample sizes within the chronic registry: chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), and prostate cancer (Supplementary Table S4). For each condition, we constructed a 1:1 case-control cohort matched by age, sex, and number of ICD-9 codes as a proxy for healthcare utilization (Supplementary Table S5, Supplementary Figure S5). Using the 5 years of clinical history before disease diagnoses, we computed a relative risk score for each condition and quantified patients' disease risk according to their percentile of the total score distribution.

For a target disease (here represented by a phecode embedding), we identify clinical code embeddings (features) most relevant to the disease by selecting all codes within a given radius in the embedding space (Methods). Each patient's score is computed as a weighted sum of these features, and the weight is calculated as the cosine similarity between the feature embedding and target disease embedding. If a patient does not have a given element in their record, the feature weight is set to 0. In practice, the size of the radius (k-neighbors) can be selected by comparing a validation patient dataset, similar to choosing the optimal performing p-value threshold in PRS[105]. This approach eliminates the need to pre-specify a list of features useful for diseases with heterogeneous presentations and highly variable symptoms, such as many chronic conditions.

Across all three conditions, we observe the separation of patients according to disease status, particularly at the extreme end of the risk distribution (Figure 4A-C). Within the group of individuals in the 90th percentile of the risk score distribution for CKD, there is a disease prevalence of 56.7%. For patients above the 98th percentile, the disease prevalence increases to 64.6%. Top-identified clinical code embeddings are consistent with each disease's known comorbidities, lab test results, and medication usage. For example, among the top features of chronic kidney disease is *Diabetes mellitus*, which is a well-documented risk factor for the disease, and *Proteinuria*, which can be an early disease indicator. Among the top features for the prostate cancer risk score are *Dysuria* and *Urinary retention*, which are both common symptoms of the condition. Although we assess the population-level correlation between disease prevalence and risk score percentile, these results suggest opportunities to leverage clinical knowledge embeddings for patient monitoring in a way that is independent of any training dataset.

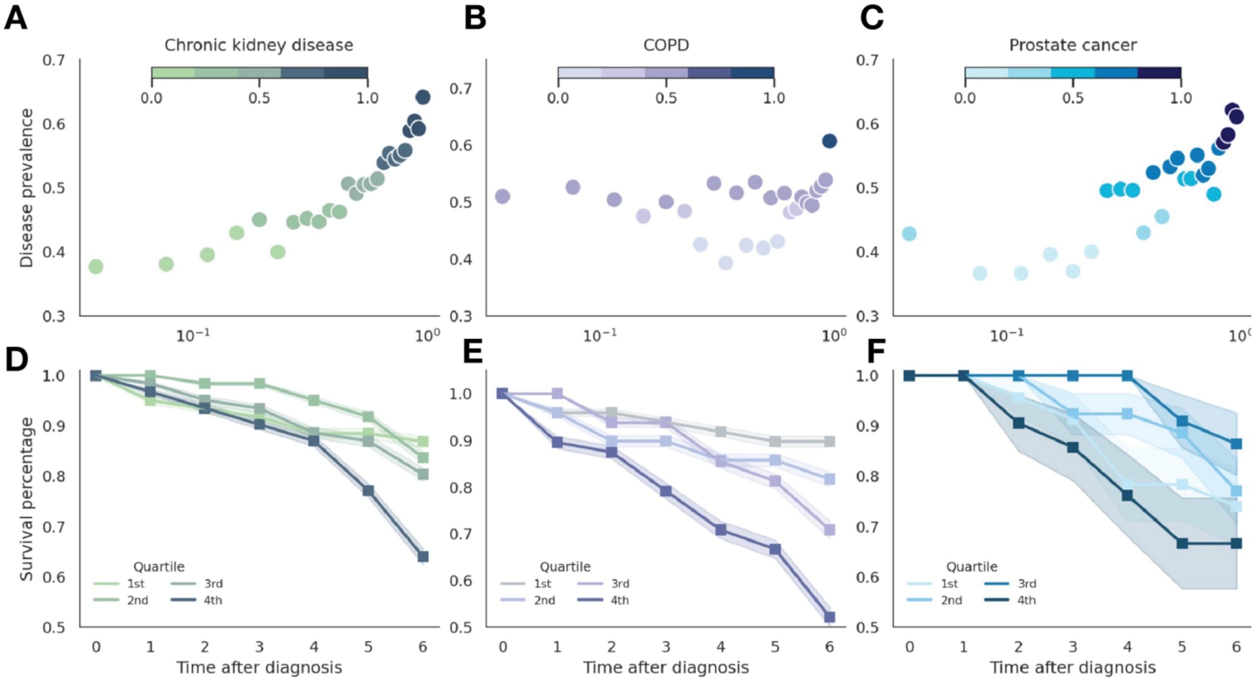


Figure 4: Clinical code embedding risk scores correlate with disease prevalence and severity. (A-C) In the top row, we provide scatter plots showing the risk score percentiles of each disease versus the disease prevalence within each bin. Dots are shaded according to disease prevalence. **(D-F)** The bottom shows a survival curve of cases that match the most common sex and age at diagnosis for each disease cohort. Lines are shaded by score quartiles. Standard error intervals are computed using 10,000 bootstrap samples. Columns are divided by diseases: chronic kidney disease, COPD, and prostate cancer.

Given the marked stratification by disease status, we evaluate whether the computed risk scores could provide disease subtyping information. Intuitively, suppose individuals acquire a greater number of highly relevant clinical conditions. In that case, this trend may indicate a later disease stage and increased mortality rates. To prevent confounding from disease-specific effects with age and sex, we restrict the assessment to individuals of the same sex and diagnosed at a similar age (Methods). For each of the three conditions, we compute the mode age and sex and use this as inclusion criteria for the survival analysis. Re-calibrating score percentiles based on the final set of patients, we observe a striking stratification by 6-year survival rate (Figure 4D, E, F). For CKD (age=69, sex=Male), individuals within the top quintile of this score distribution for CKD had a relative 28% lower survival compared to those in the bottom quartile (top-quartile: 60.3%, *bottom-quartile*: 84.1%). This trend is particularly remarkable considering the risk scores were inferred without any institution-specific, disease-specific, or patient-level information during training.

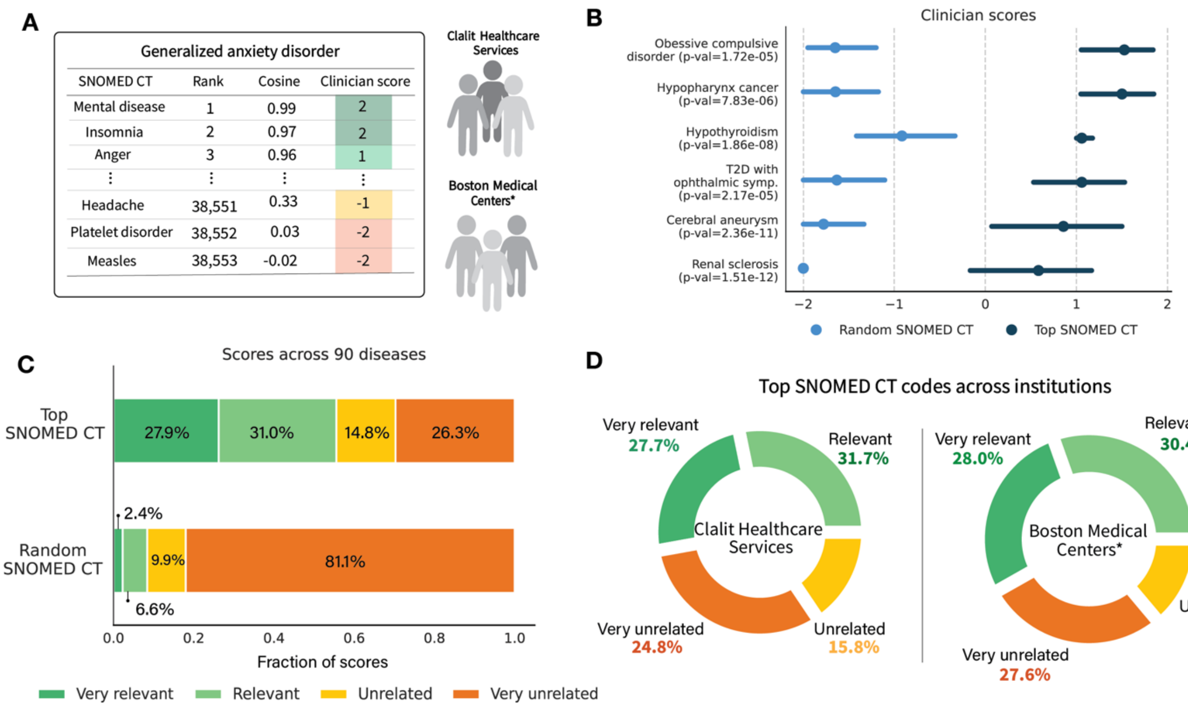
Clinical vocabulary embeddings capture medical knowledge consensus across institutions

For clinical knowledge embeddings to effectively incorporate external domain knowledge into machine learning models, the learned representations must be consistent with the current scientific consensus and reflective of the information used in regular healthcare practices. To assess this, we perform a comprehensive human evaluation study comprising expert clinicians from Israel and Boston, United States—each home to multiple leading academic and scientific research centers. We asked the panel to evaluate disease-symptom relationships inferred from the learned embedding space to quantify the alignment between information captured by the clinical knowledge embeddings and the scientific consensus. The expert panel was instructed to consider accuracy and relevance, where relationships must be accurate and relevant information likely used in typical clinical practice.

We base the evaluation study on the relationships between disease phenotypes and selected SNOMED CT codes identified within the embedding space. We specifically chose to focus on SNOMED CT codes for evaluating clinical knowledge consensus, given their utility in knowledge-driven clinical care and management resources such as clinical support decision systems and public health databases[106]. Based on expert input from a subset of clinicians, we selected 90 diseases, in the form of phecodes, that were identified as clinically relevant to the general adult patient population. The final diseases include multiple chronic diseases, neoplasms, and other common conditions (Supplementary Table S6). We selected the top 20 SNOMED CT codes for each disease based on the cosine similarity score with the target phecode embedding and 20 randomly selected SNOMED CT codes (Figure 5A). To prevent biases due to specific or uncommon symptoms, we match each top-selected SNOMED CT code with a random SNOMED CT code with a similar node degree within the KG (Methods). This process prevents diseases with prevalent symptoms from being matched with highly specific SNOMED CT codes, which could overestimate performance. Codes already forming an edge with the disease node within the knowledge graph were excluded from the analysis. The final evaluation dataset comprises 90 disease lists, each with 40 SNOMED CT codes. Clinicians were randomly assigned a set of disease lists, and each was scored once.

To quantify the relevance of each code with a disease, clinicians were instructed to grade each SNOMED CT code on a scale from -2 (unrelated) to 2 (very relevant) (Figure 5B). Across all diseases, 58.9% of top selected SNOMED CT was graded as ‘very relevant’ or ‘relevant,’ compared to only 9.0% of the randomly selected codes, providing evidence that relevant SNOMED CT codes are proximally located within the embedding space (Figure 5C). Performing a two-sided t-test for each of the 90 diseases, the average p-value is 1.02×10^{-4}, indicating a significant difference in the score distributions of the groups even after adjusting for multiple tests. Although the top SNOMED CT codes did not have a connection to the target disease in the clinical KG dataset, the embedding space still captures relationships between clinical concepts. For example, ‘major depression’ and ‘affective disorder’ are among the codes with the highest scores for obsessive-compulsive disorder (OCD). Given the high comorbidity rate and overlap of symptoms between affective disorders and OCD, these conditions are represented as nearby clinical codes in the embedding space[107]. This proximity reflects the clinical reality of their frequent co-occurrence and shared symptomatology. This analysis provides strong evidence of alignment between the disease patterns captured by clinical code embeddings and clinical knowledge. This congruence demonstrates that these embeddings can accurately represent complex relationships among disorders, potentially enhancing the understanding and prediction of comorbidities.

One advantage of unified clinical vocabulary embeddings is their ease of use across institutions. By unifying seven distinct medical vocabularies, these embeddings provide a standardized representation of clinical knowledge, bridging gaps between disparate coding systems. Importantly, they are entirely free of patient-level information, ensuring no risk of compromising patient privacy. This design makes them highly portable and suitable for deployment in diverse healthcare settings. Moreover, the embeddings are generated to capture generalizable clinical knowledge that aligns with universal medical practices, enabling consistent application across institutions with varying healthcare infrastructures. To evaluate this, we compare the distributions of scores graded by clinicians based in Israel, specifically CHS, and those graded by clinicians based in Boston, specifically Boston Children's Hospital, Harvard Medical School, and Massachusetts General Hospital, collectively referred to as 'Boston Medical Centers' throughout this work. Comparing the distributions of scored SNOMED CT codes from CHS and those from the Boston Medical Centers, we find a remarkably high level of consistency (Figure 5D). For example, for the top SNOMED CT codes, clinicians from CHS labeled 59.4% of codes as either 'highly relevant' or 'relevant'; this aligns with the score distribution from the Boston clinicians who labeled 58.4% of codes within these categories. The agreement in scoring distributions is striking, considering different healthcare systems and variations in medical training and clinical practices across countries. By integrating knowledge from multiple vocabularies and reflecting shared clinical standards, our unified clinical vocabulary embeddings offer a robust tool for facilitating interoperability and fostering collaboration in research and clinical care without the constraints of localized data dependencies. These features position embeddings as a powerful resource for advancing clinical AI in a secure, scalable, and universally applicable manner.



Generalized anxiety disorder

SNOMED CT	Rank	Cosine	Clinician score
Mental disease	1	0.99	2
Insomnia	2	0.97	2
Anger	3	0.96	1
⋮	⋮	⋮	⋮
Headache	38,551	0.33	-1
Platelet disorder	38,552	0.03	-2
Measles	38,553	-0.02	-2

Figure 5: Human pilot evaluation of clinical vocabulary embeddings shows consensus with medical knowledge across institutions. (A) An example of SNOMED CT codes ranked similarly to the phecode embedding for generalized anxiety disorder. Clinicians provide a score from -2 (least) to +2 (most), denoting each code's relevance to the disease of interest. **(B)** Clinician score distributions comparing the top-ranked SNOMED CT codes and a randomly selected set of codes across six diseases. **(C)** The distribution of clinician scores across 90 diseases. **(D)** Comparison of score distributions stratified by clinician location. * 'Boston Medical Centers' collectively refers to clinicians from Boston Children's Hospital, Harvard Medical School, and Massachusetts General Hospital.

Discussion

This work presents a unified embedding resource composed of 67,124 clinical vocabulary embeddings, providing a unified map of clinical knowledge for machine learning integration. These embeddings capture patterns between clinical codes consistent with human anatomy and clinical healthcare organization. They capture semantic properties that mirror how clinical concepts can be meaningfully combined and compared. We demonstrate that these clinical embeddings enable patient-level predictions across diverse conditions—without requiring patient data during training. Evaluations of clinical embeddings by inter-institutional panels of clinicians from the United States and Israel show that the clinical knowledge embeddings are consistent with the broader scientific and clinical research. This resource provides a safe way to share clinical knowledge and can support the development of more generalizable and transferable prediction models.

The importance of clinical expertise in developing AI models is widely recognized, and various strategies to integrate specialized knowledge have been developed[108–110]. Previous approaches rely on a single clinical ontology (often ICD-9 or ICD-10), but this overlooks the inherent heterogeneous nature of EHRs[111–115]. Furthermore, these frameworks often leverage shallow embedding techniques, which only retain local neighborhood information and struggle to incorporate heterogeneous data sources[77,116–121]. Recent techniques allow for the inference of deep embeddings, which can capture higher-order relationships. However, these often require patient-level training data and a pre-defined prediction task objective[122–124], making the resulting representations inherently dependent on the chosen patient dataset. In contrast, our approach leverages a comprehensive clinical knowledge graph that integrates relationships across multiple clinical vocabularies and a graph neural network architecture to produce highly contextualized embeddings. These embeddings capture local and global structures within the graph, providing a nuanced representation of clinical knowledge[125].

Clinical code embeddings provide a principled mechanism for representing clinical knowledge in AI models. By integrating a network of relationships across clinical codes, the embeddings provide knowledge-grounded representations of clinical data, which can serve as the basis for AI foundation models. Foundation models are large-scale models pre-trained on vast amounts of unlabeled data, which can then be adapted to tasks using smaller labeled datasets for subsequent fine-tuning steps[126]. The foundation model paradigm has been effective in natural language processing and computer vision, partly due to the open sharing of these large pre-trained models, such as GPT[127] and BERT[128]. Our resource could also be leveraged in conjunction with techniques in retrieval-augmented generation (RAG)[129] where the embeddings serve as a knowledge library. By incorporating clinical knowledge embeddings, RAG systems can retrieve information relevant to the user prompt and contextually aligned clinical vocabulary embeddings.

The clinical knowledge embeddings are a shareable pre-trained model that can be transferred across institutions without risking patient information leakage. Unlike models pre-trained on individual-level patient records, where there is no guarantee that the model parameters will not inadvertently retain or memorize private information[130,131], our unified clinical code embeddings can be securely exchanged between healthcare systems without risk of patient data exposure. This approach could also reduce sensitivity to variations in patient-level data, allowing for more consistent predictions across diverse patient populations. A cross-institutional analysis in Israel and the United States indicates that these embeddings capture universal clinical patterns. By providing a standardized, privacy-preserving representation of clinical knowledge, these embeddings can facilitate collaboration and model development across healthcare systems and geographical regions, potentially accelerating advances in medical AI.

While this resource holds potential for advancing precision medicine, several limitations exist. First, the current knowledge representations cannot capture varying levels of uncertainty. This is especially relevant since clinical knowledge is compiled from numerous sources, and the frequency of specific clinical observations can often vary. For instance, disease guidelines may list multiple symptoms, but the frequency and prominence of these symptoms can vary widely across individuals. Additionally, the quality of the embeddings is highly dependent on the underlying knowledge graph. Inaccuracies, such as incorrectly defined edges, can introduce noise and distort the flow of information through the graph. As clinical knowledge evolves, it is paramount that the knowledge graph is continually updated to reflect the current clinical consensus. Another limitation is the absence of contextualizing information, such as age and sex, that can impact the accuracy and relevance of clinical knowledge elements. For example, the clinical presentation of a heart attack can vary significantly between males and females[132], but this delineation is not considered in the construction of the current embeddings.

Clinical vocabulary embeddings represent common and chronic diseases and offer limited coverage of clinical concepts related to genetic disorders and Mendelian diseases. In future work, we aim to address this gap by incorporating codes from phecodeX[133](a more recent version that includes chromosomal anomalies) and knowledge bases specifically developed to document Mendelian disorders, such as the Online Mendelian Inheritance in Man (OMIM)[134] and Human Phenotype Ontology (HPO)[73]. We also plan to integrate databases that capture biological mechanisms, including those related to gene and protein processes, which could enhance the characterization of diseases with complex clinical presentations and poorly defined conditions.

This unified clinical vocabulary embedding resource transcends institutional boundaries by integrating seven medical vocabularies into a cohesive, standardized representation. This unification provides a robust foundation of clinical knowledge, eliminating the challenges posed by fragmented coding systems while safeguarding patient privacy. Designed to capture generalizable clinical knowledge, these embeddings align with medical practices. Unlike datasets limited to health state measurements, these embeddings incorporate a wealth of clinical knowledge, including mechanistic rationales, clinical guidelines, and disease and treatment pathways. By embedding this knowledge into AI, the resource can help make AI models evidence-based and clinically grounded[135]. Ensuring that AI models reflect clinical knowledge and resources, such as unified clinical vocabulary embeddings, can enhance reliable cross-institutional use of AI.

Methods

Construction of clinical knowledge graph

Standardized medical vocabularies

To provide maximum flexibility and utility of the clinical knowledge graph and subsequent embeddings, we derive all clinical concepts from standardized medical vocabularies: International Classification of Diseases (ICD)[136], Anatomical Therapeutic Chemical (ATC) Classification[83], RxNorm[84], Systemized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[106], Current Procedural Terminology (CPT)[137], Logical Observation Identifiers Names and Codes (LOINC)[138], and phecodes[139]. Here, we define a clinical code as an entry within a medical vocabulary with a unique identifier.

We begin by integrating the clinical codes and relationships defined in the phecode ontology. This database represents clinical phenotypes as phecodes, formed by aggregating groups of diagnostic codes (i.e., ICD codes) into clinical sub-groupings. In total, 13,707 ICD-9 codes are mapped to 1,817 phecodes[82,140,141]. Specifically, phecodes were developed to facilitate the secondary use of EHRs for biomedical research rather than for patient monitoring or hospital administration. A phecode is a 3-digit parent code with additional digits following the decimal point. Numbers after the decimal point reflect a hierarchical structure similar to the ICD code hierarchy. Under each leaf phecode is the set of corresponding ICD-9 codes where each phecode has between 1 – 20+ associated ICD-9 codes. We treat all parent phecodes and subsequent children phecodes (collectively referred to as ‘phecodes’ in this work) as individual clinical codes. Each ICD-9 code is also treated as a separate code, and we define a relationship between two clinical codes if they have a direct connection within the phecode hierarchy. This process provides an initial set of 15,524 clinical codes and 20,783 pairwise relationships.

Next, we incorporate the clinical codes and relationships defined by the PheMap database, an existing knowledge base of biomedical knowledge that defines a set of relationships between the codes and various clinical vocabularies[81]. This database offers relationship pairs linking phecodes and clinical codes from other medical vocabularies, including RxNorm, CPT, LOINC, and SNOMED CT. Relationships defined in PheMap were determined through text mining and natural language processing of open-source biomedical literature. These relationships can be interpreted as a phenotype being "associated with" a given clinical concept. However, we treat every relationship between clinical concepts as the same type regardless of the original database. In PheMap, clinical codes from the LOINC vocabulary are encoded as LOINC part codes, standardized attribute values used to construct full LOINC codes that specify a given laboratory test. Overall, this dataset yields a set of 78,801 codes and 727,939 pairwise relationships[142].

Finally, we utilize the Unified Medical Language System (UMLS) database to incorporate additional levels of information not provided in the previous databases, specifically information about the ATC vocabulary. Using the UMLS database, we map each RxNorm code to a Concept Unique Identifiers (CUI)[143] if the given RxNorm code is the preferred term listed for the CUI. For each CUI, we select all ATC-5 codes that fall under that particular CUI with the term type 'IN' (ingredient), meaning that the given CUI represents a concept that is an ingredient within the specified ATC-5 class. For example, the ATC-5 code for the drug 'prednisolone' would be obtained via: 8638 (RxNorm) to C0032950 (UMLS CUI) to S02BA03, D07AA03, …, C05AA04 (ATC-5). The mapping between UMLS CUI and ATC-5 is not 1:1, so CUIs may map to multiple ATC-5 codes. To establish relationships between each ATC-5 code and codes from other clinical vocabularies, we first map each clinical code in the previous PheMap database to its corresponding UMLS CUI code. Next, we create a pairwise relationship between an ATC-5 code and a phecode if the PheMap database shows a relationship between a clinical concept with the same CUI code and that phecode.

Building a heterogenous knowledge graph

Leveraging the clinical codes and relationships drawn from existing biomedical knowledge bases and vocabularies, as described above, we construct a knowledge graph[144] (KG) to efficiently organize clinical knowledge in a format optimized for machine learning. First, we designate each clinical code as an individual node and the corresponding node type as the terminology source. Next, we draw an undirected edge between two clinical codes if they have a pairwise relationship described above in the clinical vocabulary aggregation process. Multiple edges are drawn if two nodes have a relationship defined multiple times (e.g., a relationship between a phecode and

ICD-9 code derived from the phecode hierarchy and the PheMap knowledgebase). Note that all edges are treated as the same edge type.

An advantage of knowledge graphs over traditional relational databases is that KGs enable the representation of complex, multi-dimensional relationships. To accomplish this, we can incorporate additional semantic layers and hierarchies between clinical codes. First, we can leverage the natural hierarchical structure of ATC codes to incorporate medication information at varying granularities. We merge existing ATC-5 nodes according to their ATC-4 class to create a broader representation of therapeutic information as represented by ATC-4 nodes. All edges from the aggregated ATC-5 nodes are mapped to the new ATC-4 node during this process. Because the RxNorm nodes already represent medication on a drug-level granularity, we remove the ATC-5 nodes and retain the ATC-4 and RxNorm nodes for representing medications.

Next, we consolidate nodes representing similar information to reduce redundancies within the graph, enabling the sharing of neural network information and enhanced interpretability. This is especially relevant for LOINC part codes, where different types of attributes will have different codes even if the entity of interest is the same. For example, 'LP7501-2' and 'LP200001-8' have the description 'Prostate', but the first is a system LOINC part code, and the latter is a radiology LOINC part code. We also merge sets of SNOMED CT codes that have a shared synonym. Finally, CPT codes with the same description are also merged. This occurs when procedures have the same description but different identifiers due to slight variations in the procedure. Altogether, we merge 2,507 clinical concepts. Edges from the previous individual nodes are all aggregated into the newly merged node.

We filter nodes with a very high degree (number of directly connected neighbors), as high-degree nodes can lead to the overrepresentation of certain information in the resulting embeddings. Additionally, nodes with an extremely high degree present computational challenges, especially when utilized in the context of graph neural networks. We threshold nodes based on a degree of 1,000, which identifies 64 clinical codes. The majority of these are phecodes representing infectious diseases ('Tuberculosis', 'Meningitis') and symptoms ('Cough', 'Pain in joint'), and broad disease classifications ('Cerebrovascular disease', 'Other conditions of brain'). The high node degree for infectious diseases is primarily due to the detailed ICD-9 relationships describing disease strain and location of infection. We remove edges between the phecode and ICD-9 node for each high-degree node if the ICD-9 code contains >1 place after the decimal point (e.g., 005.81). This allows us to maintain broad levels of information in a node's neighborhood while discarding overly detailed information. The final graph comprises 67,124 nodes across seven node types and 1,315,610 edges, and the distribution of node types can be found in Supplementary Table S1.

Heterogenous graph transformer

Graph notation

A heterogeneous graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, is defined by a set of nodes $\mathcal{V}$ and edges $\mathcal{E}$. We define a mapping function $f_v: \mathcal{V} \to \mathcal{T}^v$where each node $v \in \mathcal{V}$ has an assigned node type $f_v(v) \in \mathcal{T}^v$. Here, the node types represent different clinical vocabularies. For a given graph, $\mathcal{G}$, we denote the graph neural network, $\mathcal{H}_{\mathcal{G}}$, as a function that maps a vector of features for a node u, written as $\boldsymbol{x_u} \in \mathbb{R}^d$, to a vector of real numbers representing the latent space:

$$\mathcal{H}_{\mathcal{G}}: \boldsymbol{x_u} \to \boldsymbol{z_u} .$$

Here, d is the dimension of the input vector. This resulting mapped vector (also called an *embedding*) is denoted by $\boldsymbol{z} \in \mathbb{R}^d$. For convenience of notation, we assume that the input and output dimensions are equal. We provide a table with relevant notation and indicate whether they are an inferred parameter within the model.

Variable	Description	Dimension	Trainable (Y/-)
$\boldsymbol{\mathcal{G}}$	Graph	-	-
$\boldsymbol{\mathcal{V}}$	Set of vertices or nodes	-	-
$\boldsymbol{\mathcal{E}}$	Set of edges	-	-
$\boldsymbol{\mathcal{T}}^{v}$	Space of node types	-	-
d	Input dimension	$\mathbb{R}^1$	-
L	Number of layers	$\mathbb{R}^1$	-
$\boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{G}}}$	Graph neural network function	$\mathbb{R}^d \rightarrow \mathbb{R}^d$	Y
$\boldsymbol{x_u}$	Feature vector of node *u*	$\mathbb{R}^d$	-
$\boldsymbol{z}$	Embedding vector of node *u*	$\mathbb{R}^d$	Y
$\boldsymbol{\mathcal{N}(u)}$	Neighborhood of node *u*	-	-
$\boldsymbol{m_u}$	Message feature of node *u*	$\mathbb{R}^d$	Y
$\boldsymbol{h}_{\boldsymbol{u}}^{(l)}$	Embedding of node *u* at the *l*th layer	$\mathbb{R}^d$	Y
$\mathbf{ATT}(\boldsymbol{s},\boldsymbol{t})$	Attention vector for node *u*	$\mathbb{R}^d$	Y
h	Number of heads	$\mathbb{R}^1$	-
$\boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{i}}, \boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{i}}$	Weight matrices	$\mathbb{R}^{d\times\frac{d}{h}}$	Y
$\boldsymbol{head}(\boldsymbol{s},\boldsymbol{t})^{\boldsymbol{i}}$	Attention head vector for source node *u* and target node *v*	d/h	Y
$\boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{i}}, \boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{i}}$	Weight matrices for attention computation	$\mathbb{R}^{d\times\frac{d}{h}}$	Y
$\boldsymbol{W}_{\boldsymbol{V}}$	Weigh matrix for attention computation	$\mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$	Y
$\boldsymbol{K}_{\boldsymbol{s}}^{\boldsymbol{i}}, \boldsymbol{Q}_{\boldsymbol{t}}^{\boldsymbol{i}}$	Query and key vectors	$\mathbb{R}^{\frac{d}{h}\times 1}$	Y
$\boldsymbol{A} \odot \boldsymbol{B}$	Hadamard product between **A** and **B**	-	-
$\bigoplus$	Element-wise summation	-	-
$\big\Vert_{\boldsymbol{n}}(\boldsymbol{x}^{\boldsymbol{i}})$	Row-level concatenation of the set vectors	-	-
$\boldsymbol{f}(\boldsymbol{u},\boldsymbol{v})$	Similarity scoring function	-	Y
$\Vert\cdot\Vert$	L2-norm	$\mathbb{R}^1$	-
$\boldsymbol{v}^{+}$	Node forming a positive edge with $\boldsymbol{u}$	-	-
$\boldsymbol{v}^{-}$	Node that does not form an edge with $\boldsymbol{u}$	-	-
$\boldsymbol{\mathcal{G}}'$	Graph with masked edges	-	-

Table 1: Summary of notation. We provide a summary of the variables used in deriving the graph transformer network model, the dimension of each variable, and whether it is a learned or fixed parameter. Rows are sorted according to the order in which each variable is introduced.

Overview

We model the architecture of our graph neural network (GNN) after the heterogeneous graph transformer (HGT)[85]. In this work, we briefly describe the key model steps. The complete derivation of HGT is detailed in previous work[85].

Learning $\mathcal{H}_{\mathcal{G}}$ involves aggregating information across the graph structure defined by $\mathcal{G}$. For a target node t, we define its neighborhood, $\mathcal{N}(t)$, as the set of nodes with a direct connection or edge with *t*. Information is propagated through the graph by gathering information from the neighborhood of t, which is aggregated into a message feature $\boldsymbol{m}_t$. The embedding of *t* can be updated with the resulting message feature, $\boldsymbol{h}_u^{(l+1)} = \mathrm{UPD}\left(\boldsymbol{m}_t^{(l)}, \boldsymbol{h}_{\boldsymbol{u}}^{(l)}\right)$ where $\boldsymbol{h}_{\boldsymbol{t}}^{(\boldsymbol{l}+\mathbf{1})}$is the embedding of node t at the *l+1* layer and UPD$(\cdot)$ refers to an update function that integrates messages with the current node embedding. Note that the input of the *l+1* layer is the output of the previous layer, and by stacking L layers $\boldsymbol{h}_t^{(\cdot)}$, we represent the final HGT as a composition of these functions across all layers:

$$\mathcal{H}_{\mathcal{G}}(\boldsymbol{x}_t) = \left(\boldsymbol{h}_t^{(L)} \circ \boldsymbol{h}_t^{(L-1)} \circ \ldots \circ \boldsymbol{h}_t^{(1)}\right)(\boldsymbol{x}_t).$$

The input of the model is the set of node features, $\boldsymbol{x}_t \in \mathbb{R}^d$ for all nodes in $\mathcal{G}$ which is then used to initialize the 0^{th} layer embedding of each node $\boldsymbol{h}_t^{(0)}$. In practice, each of the node features is initialized using Xavier noise. The embedding represents the relational information learned from the knowledge graph without additional information from the node features.

Mutual attention

The HGT utilizes the idea of multi-head attention, which is a key feature of the traditional transformer architecture[86]. Multi-head attention utilizes multiple attention mechanisms, or heads, in parallel to tend to different parts of the input, where each head has its own set of learnable weights (parameters). Updating a node's embedding involves aggregating messages across all h heads and weighting each message by its attention score. This mechanism essentially up- and down-weights messages based on their importance so that only necessary information is propagated to the target node. Although the original HGT architecture allows the modeling of different relation types within heterogeneous graphs, we treat all nodes and edges as a single type during inference to maximize information sharing across the knowledge graph.

For a target node t, a single attention head is a series of linear projections of the embedding from the previous layer, normalized by the embedding dimension. Here, a linear projection for an input $\boldsymbol{x} \in \mathbb{R}^d$, is defined as an affine linear transformation, $\boldsymbol{y} = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{b}$ is a bias term and $\boldsymbol{W}$ is a learnable weight matrix. For convenience, we drop the bias term in the following derivations, but in practice, the parameter corresponding to the bias term is absorbed into the estimation of $\boldsymbol{W}$.

The series of linear projections mirror the standard query, key, and value vectors used in the standard transformer architecture. The attention score vector corresponding to the edge between source node s and target node t, represented as $\mathbf{ATT}(s,t)$, is computed as the concatenation of each head's output:

$$\boldsymbol{K}_s^i = \left(\boldsymbol{W}_K^i\right)^T \boldsymbol{h}_s^{(l-1)},$$

$$\boldsymbol{Q}_t^i = \left(\boldsymbol{W}_Q^i\right)^T \boldsymbol{h}_t^{(l-1)},$$

$$\boldsymbol{HEAD}(s,t)^i = \left(\boldsymbol{W}_V^i\right)^T \boldsymbol{K}_{\boldsymbol{s}}^{\boldsymbol{i}} \odot (\boldsymbol{Q}_{\boldsymbol{t}}^{\boldsymbol{i}}) \cdot \frac{1}{d},$$

$$\mathbf{ATT}(s,t) = \Big\|_{i\in[0,h]} \mathbf{HEAD}(s,t)^i.$$

Here, the operator "$\boldsymbol{A} \odot \boldsymbol{B}$" represents the elementwise or Hadamard product between $\boldsymbol{A}$ and $\boldsymbol{B}$, and the operator "$\|_n(\boldsymbol{x}^i)$" represents the row-level concatenation of the given set of vectors $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^n\}$.

The input vector is $\boldsymbol{h}_s^{(l-1)} \in \mathbb{R}^{\boldsymbol{d}\times\mathbf{1}}$, and the learnable weight matrices are $\boldsymbol{W}_Q^i, \boldsymbol{W}_K^i \in \mathbb{R}^{d\times\frac{d}{h}}$, and $\boldsymbol{W}_V \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$ where $\frac{d}{h}$ is the input dimension divided by the number of heads. For these weight matrices, the subscripts refer to the *query*, *key*, and *value* elements in the transformer. The resulting query and key vectors follow as $\boldsymbol{K}_s^i, \boldsymbol{Q}_t^i \in \mathbb{R}^{\frac{d}{h}\times 1}$. The resulting attention vectors from each head and final multi-head attention vector are $\mathbf{HEAD}(s,t)^i \in \mathbb{R}^{\frac{d}{h}\times 1}$ and $\mathbf{ATT}(s,t) \in \mathbb{R}^{d\times 1}$. And to ensure the attention scores for the target node *t* sum to 1, a $Softmax(\cdot)$ is applied across all neighbors of *t*.

Message passing

For target node *t,* we first compute the message vectors across all heads for each neighbor of *t*. This message is calculated as a linear projection of the embedding for source node *s* from the prior layer multiplied by a weight matrix ($\boldsymbol{W_T}$). Each head maintains its linear projection matrix but $\boldsymbol{W_T}$ is shared across all attention heads.

The final message passed from $s\ \rightarrow t$ is a concatenation of messages produced from each of the *h* heads. This resulting message vector is then multiplied by the multi-head attention vector. The final updated embedding ($\boldsymbol{h}_t^{(l)}$) is computed by aggregating across all neighbors of the node t through an elementwise sum. This is followed by a linear projection ($\boldsymbol{W}_N$), which creates separate projection for each node type. We need a single matrix because we only model a single node type. This is then followed by a residual connection and passed through an activation function:

$$\boldsymbol{m}_s^{(l)} = \big\|_{i\in[1,h]}\left(\boldsymbol{W_T}\,(\,\boldsymbol{W}_M^{i^T}\boldsymbol{h}_s^{(l-1)})\right),$$

$$\widetilde{\boldsymbol{h}}_t^{(l)} = \bigoplus_{\forall v\in N(u)} \mathbf{ATT}(s,t)\odot \boldsymbol{m}_s^{(l)},$$

$$\boldsymbol{h}_t^{(l)} = \sigma\left(\boldsymbol{W}_N\,\widetilde{\boldsymbol{h}}_t^{(l)} + \boldsymbol{h}_t^{(l-1)}\right).$$

The learnable weight matrices are $\boldsymbol{W_T} \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$, $\boldsymbol{W}_M^i \in \mathbb{R}^{d\times\frac{d}{h}}$, and $\boldsymbol{W}_N \in \mathbb{R}^{d\times d}$. The message vector is $\boldsymbol{m}_t^{(l)} \in \mathbb{R}^{d\times 1}$. The operator $\oplus$ represents the elementwise summation. Training the final model involves optimizing the set of learnable weight matrices, $\{\boldsymbol{W}_K^i, \boldsymbol{W}_Q^i, \boldsymbol{W}_V^i, \boldsymbol{W}_M^i\}$ across *h* heads and finding $\{\boldsymbol{W}_T, \boldsymbol{W}_N\}$ that minimizes the loss function $\mathcal{L}$.

Self-supervised learning

For a given node u, our goal is to generate an embedding $\boldsymbol{z}_u$that quantifies the topological information of the graph w.r.t u. The learning objective can be framed as a link prediction problem, where given two nodes u and v, we want to train a function to predict whether an edge exists between these two nodes. This function can be described by,

$$f(u,v) = \frac{(\boldsymbol{z}_u)^T \boldsymbol{z}_v}{\parallel \boldsymbol{z}_u \parallel \parallel \boldsymbol{z}_v \parallel} = cos(\boldsymbol{z}_u, \boldsymbol{z}_v),$$

where $\boldsymbol{z}_u = \mathcal{H}(u)$, and $\mathcal{H}(\cdot)$ represents the trained GNN function that provides the node embeddings. Note that because edges are undirected and the scoring function is symmetric, $f(u,v) = f(v,u)$.

Optimizing this above scoring functions means solving for the parameters of $\mathcal{H}(\cdot)$. This can be performed using contrastive learning. We mask a random subset of edges to create a modified version of the graph, $\mathcal{G}'$. Then, given a source node u, we define a node where $v^+ \in \mathcal{N}(u)$ represents a target node that forms an edge with u in the original graph $\mathcal{G}$. We denote $v^- \notin \mathcal{N}(u)$ as a node that does not form an edge with u in $\mathcal{G}$ and denote the set of these negative examples as $\{v_1^-, v_2^-, \cdots, v_n^-\}$.

The function $f(u,v)$ is trained on $\mathcal{G}'$to estimate whether a source node u forms an edge with target node v in the original graph $\mathcal{G}$. We can quantify the probability of an edge existing between u and v^+ with negative edges $v_{\{1:n\}}^-$ as the probability,

$$p\left(u, v^+, v_{\{1:n\}}^-\right) = \frac{exp(\boldsymbol{z}_u^T \boldsymbol{z}^+)}{exp(\boldsymbol{z}_u^T \boldsymbol{z}^+ + \boldsymbol{z}_u^T \boldsymbol{z}_1^- + \cdots + \boldsymbol{z}_u^T \boldsymbol{z}_n^-)}.$$

Given a dataset of N source nodes, each with a corresponding set of candidate nodes as described above, the negative log-likelihood over the dataset is equivalent to the cross-entropy loss where the true "class" always corresponds to the positive node v^+. This type of contrastive loss function is well-established[88,145–147] and is formulated as follows:

$$\mathcal{L} = -\sum_u^N log\left(p\left(u, v^+, v_{\{1:n\}}^-\right)\right),$$

where $v^+ \in \mathcal{N}(u)$ and $v_{\{1:n\}}^- \notin \mathcal{N}(u)$.

Node type-aware sampling of positive and negative edges

The contrastive learning requires specifying positive and negative edges during training, where a positive edge is defined as a pair of nodes with an existing edge in the KG, and a negative edge is a pair of nodes that are not connected by an edge. To generate a training tuple comprised of a single positive edge and a set of corresponding negative edges, we randomly sample an existing edge within the KG denoted as (u, v) and $\mathcal{T}^V(v)$ is the node type of target node v. To create a negative sample, we select a node r such that $\mathcal{T}^V(r) = \mathcal{T}^V(v)$ and $r \notin \mathcal{N}(u)$.

Stochastic mini-batch training with neighbor sampling

Because the topology and size of the GNN are determined by the KG, storing the entire network and its hidden states onto a single GPU is impossible. Although training with mini-batches alleviates some of this issue, the computation graph (the set of nodes involved in message passing in each iteration) for a GNN with L layers often still includes millions of parameters. To circumvent this, we use node neighbor sampling to select only a subset of nodes to include at each layer when computing a node's message vector. Specifically, we take a sample n of neighbors at each k-hop neighborhood at each gradient descent step when computing messages during optimization. Node types with a smaller set of nodes in the KG are upweighted to ensure they are represented in the sampling. The number of samples at each layer is a hyperparameter selected through hyperparameter tuning. The final training was performed on an H100 GPU for ten epochs, during which we trained on all 1,315,610 edges of the KG.

Hyperparameter tuning

We perform hyperparameter tuning in 2 steps: first for the model architecture and then for the parameters involved in the contrastive learning objective. We select a random subset of 500,000 edges for training and 100,000 for validation from the KG. We perform hyperparameter tuning using Raytune where we fix the learning rate (1e-4), number of negative samples (10), batch size (100), and sampling neighborhood (10) and vary the following parameters: number of layers (2, 3, 4), input feature dimension (128, 256, 512), output embedding dimension (128, 256, 512), head dimension (128, 256, 512), number of heads (2, 3, 4), and dropout (0, 0.1, 0.2). Performance is assessed by measuring the prediction accuracy on the set of 100,000 validation edges and 100,000 node pairs that are not linked in the KG (negative edges). The optimal set of GNN parameters is 4 layers, 128 input dimensions, 128 output embedding dimensions, 512-dimension head size, three heads, and a dropout of 0.0.

We repeat the hyperparameter selection process by fixing the parameters selected above and vary the following: learning rate (1e-3, 1e-4, 1e-5), neighborhood sampling (5, 10, 20), and number of negative samples (3, 5, 10). The final hyperparameter set selected is a learning rate of 1e-5, neighborhood sampling of 20 nodes, and three negative samples. All experiments were performed using the Ray Tune software (https://docs.ray.io/en/latest/tune/index.html).

Visualizing the latent space

To visualize the embedding space, we perform UMAP[148] directly on the embedding vectors and plot the first two dimensions. When performing dimensionality reduction, we use phecode, ATC-4, LOINC, and CPT embeddings. Phecode categorizations are pre-defined from the phecode v1.1 mapping[149]. To visualize the latent space of phecodes within the digestive system, we performed UMAP only on the subset of phecode embeddings within the 'Digestive' category. We highlight phecodes broadly corresponding to mouth/teeth conditions (phecodes 520-530), functional digestive disorders (560-570), and biliary tract disorders (570-580). To highlight embeddings related to specific diseases, we compute the cosine similarity between a phecode embedding (representing a disease) and each embedding within the latent space. We visualize this relationship by shading each embedding within the UMAP space according to the cosine similarity. We highlight relevant procedures (CPT) and laboratory tests (LOINC) for lupus and Graves' disease to provide additional interpretation of the latent space map. These were derived from the guidelines provided by the National Resource Center on Lupus[90] and the European Thyroid Association Guideline[89].

Phenotype knowledgebase comparison

To quantify the quality of the embedding space, we wanted to measure the distance or similarity between clinical concepts that have well-defined relationships externally of those defined in the knowledge graph. We first assessed the cosine similarity between a given disease (in the form of a phecode embedding) and the EHR-based codes that reflect this phenotype. For this, we utilized the Phenotype Knowledgebase (PheKB)[92], a database of validated electronic algorithms to identify patient characteristics and diseases from clinical data. Many algorithms listed in PheKB use ICD codes as a critical component for phenotyping, including those describing anxiety[150], chronic kidney disease[151], and breast cancer[152]. For each of these three diseases, we computed the cosine similarity between the corresponding disease phecode, and the list of ICD codes used in each algorithm. We compare this with a randomly selected sample of ICD codes from the embedding space and measure significance using a two-sided Mann–Whitney U test. We apply a multiple hypothesis testing correction to account for three tests.

Disease embedding arithmetic

Constructing disease symptom embeddings

To demonstrate the syntactic and semantic properties of the embedding space, we use vector-level arithmetic to aggregate together disease symptoms to form a ‘disease symptom embedding’. Aggregation is performed using sum pooling, but any vector-level aggregation method could be used in practice. For a given disease, we retrieve the list of symptoms from the Mayo Clinic Disease Description Guidelines[153]. A clinician manually translates each symptom into its corresponding ICD-9 code. We provide these translations for all nine diseases discussed within this analysis (Supplementary Table S2). To demonstrate the trend of increased similarity between a disease embedding and the disease symptom embedding, we order symptoms from highest to lowest prevalence. We approximate population prevalence using a random sample of 200,000 patients from the CHS database.

Disease symptom embeddings are first initialized with a vector of zeros. We then add the embeddings from all symptoms (represented by an ICD-9 embedding). For visualization, we compute the cosine similarity between the disease symptom embedding at each step and the target disease embedding, represented using a phecode embedding. Across diseases, we see a broad trend of increased similarity with the disease embedding as more symptoms are aggregated.

Robustness analysis

To assess the robustness and generalizability of these trends, we perform the same analysis under various conditions while perturbing different variables in the framework. This includes analyses assessing the uncertainty in selecting disease symptoms, incorrect disease specification, and the impact of random symptoms. First, due to the noise and uncertainty associated with using diagnosis codes as proxies for disease symptoms[13], we tried to model real-world circumstances by varying the ICD-9 code used to represent each symptom. We selected a set of candidate ICD-9 codes for each symptom by taking the original ICD-9 code from the clinician review and selecting all ICD-9 codes with the same parent ICD code by removing one decimal place. For example, for code '786.2' (Cough), the candidate set would include '786.1' (Stridor), '786.0' (Dyspnea and respiratory abnormalities), etc. Then, when selecting each symptom for aggregation, a random code from the candidate set is selected. This process is repeated 10 times, yielding similarity curves constructed from 10 different disease symptoms.

Next, we wanted to simulate when the disease of interest is mis-specified compared to the symptom list. This reflects situations where the symptoms do not describe the target disease of interest but, instead, a closely related condition. To imitate this, we selected a set of diseases similar to the target disease; this was done by looking at other phecodes with the same parent phecode. When computing the similarity between symptom embedding and disease embedding composition, we used the phecode embedding of the alternative disease. Finally, we wanted to assess the situation where incorrect information is included in the symptom list. We simulated this by adding a randomly selected ICD-9 code to the symptom list and included it in the computation of the aggregated symptoms. Because we use vector addition to aggregate symptoms and addition is commutative, we place the random symptom at the beginning of the aggregation procedure.

Phenotype risk score

Case/control construction

For each disease, we first select the corresponding range of phecodes for the specific disease. This complete list of phecode inclusion ranges is provided in Supplementary Table S4. Then, all ICD-9 codes falling under each phecode make up the inclusion ICD-9 codes. To define cases, we first selected all patients with at least one ICD-9 code from the corresponding inclusion codes. We then filter patients to include only patients listed on the CHS chronic registry to confirm they have received the disease diagnosis. Because patients are not guaranteed to be added to the chronic registry when they are first diagnosed, we treat the first occurrence of the inclusion ICD-9 code as the date of diagnosis. For each patient, we set the index date as January 1 of the year of diagnosis, where only information occurring before the index date will be included in the risk score computation. For prediction, we select all EHRs recorded within the past 5 years of the index date.

To define controls, we begin with the phecode exclusion criteria provided by the phecode mapping v1.1. For each phecode, the mapping provides a list of exclusion phecodes that are typically similar conditions or overlap with the target disease. We begin by selecting all individuals who have never received any exclusion diagnoses at any point in their EHR. We also remove individuals on the chronic disease registry for the specified disease. We select a matched control from this set of candidate control patients for each case. Matching is based on the following criteria: 1) same sex, 2) year of birth ±2 years, and 3) number of recorded diagnoses in the EHR (with at least 1 diagnosis in the EHR). The complete case-control process is illustrated in Supplementary Figure S5.

Risk score computation

For a given disease, a risk score is computed as the weighted sum of relevant clinical codes, where the weight is the cosine similarity between the embedding of each code and the disease embedding. We restricted the set of potential features to laboratory tests, medications, and ICD-9 codes with at least a 0.1% prevalence within the CHS patient population. The CHS EHR did not have SNOMED CT or CPT codes for analysis. To determine the clinical codes related to each disease, we compute the cosine similarity between all clinical codes within the latent space and the target disease embedding, represented by a phecode. The final set of features is the top k codes, where k=150, and the feature weights are the cosine similarity scores. In practice, the choice of k can be treated as a hyper-parameter and optimized with a validation dataset, similar to the p-value thresholding of polygenic risk scores. Patient risk scores are computed as a weighted sum where the associated weight is added to the sum if the patient has the given code in their EHR and given a 0 otherwise. Features are counted once in the score computation, even if a patient receives the codes multiple times.

Survival analysis

For each disease, we computed the most common age at diagnosis and the sex with the highest disease prevalence. We restricted the survival analysis to patients meeting these age/sex criteria and those with a diagnosis date from 2016 or earlier. We re-computed risk score quantiles based only on this restricted set of patients. We looked at the 6 years of EHR after each patient's index date and denoted a patient as deceased if they had a non-null death date. Patients were stratified into risk score quartiles based on the risk score computed at the index date. Confidence intervals for the survival curves were computed using 10,000 bootstrapped samples.

Clinical evaluation study

A key transferability aspect is demonstrating that the inferred clinical knowledge embeddings capture medical insights that align with established clinical consensus. Although the knowledge graph construction was based on mined biomedical literature, the inferred embeddings and latent space offer insight into relationships between codes beyond the direct edges in the KG. To validate this, we perform a human pilot evaluation with groups of clinical experts from Brigham Women's (N=1), Mass General (N=1), Boston Children's Hospital (N=1), and CHS (N=3).

To systematically assess this, we evaluate the association between different diseases and SNOMEDCT US codes. We restrict the set of SNOMED codes to the Clinical Observations Recordings and Encoding (CORE) Problem List Subset of SNOMED CT[48]. This subset of codes has been identified as most informative in documenting and encoding clinical information. We survey clinicians to select 90 disease phenotypes (phecodes) that broadly capture major disease areas and are clinically relevant for potential clinical AI applications. For each disease, we compute the cosine similarity between each SNOMED CT code that does not already form an edge with the target phecode. We select the top 20 SNOMED CT codes for evaluation. A control set is constructed by randomly selecting 20 SNOMED CT codes that are not in the top 20 nor already form an edge in the KG. To prevent biases due to differences in the frequency of symptoms or high degree and low degree within the KG, we match control SNOMED CT codes based on the case code's node degree.

Clinicians were provided with shuffled disease lists of case and control codes. For a given disease, a single clinician evaluated the full list of codes. They were provided a brief background of SNOMED CT codes and a scoring rubric where they were asked to evaluate each SNOMED CT code according to how related each code is to the given disease phenotype. We utilize a Likert scale[154] based on scores -2 – +2: ‘unrelated’ (-2), ‘unsure’ (0), and ' highly related’(+2). If a code is too broad or non-specific (e.g. ‘Clinical evaluation’), these are assigned a 0.

Institutional Review Board approval

Parts of this study that relate to the use of CHS data (for evaluating the predictive ability of the KG-derived concept embeddings) were approved by the CHS Institutional Review Board (Helsinki) committee.

Data availability

Data and clinical concept embeddings, as well as the PheKG knowledge graph, are available via Harvard Dataverse at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Z6H1A8. Due to national and organizational data privacy regulations, CHS individual-level data from this study cannot be shared publicly.

Code availability

The Python implementation of the heterogeneous graph transformer and the associated codebase for model training, (non-individual-level) analyses, and embedding evaluations are available at https://github.com/mims-harvard/Clinical-knowledge-embeddings. The project website is at https://zitniklab.hms.harvard.edu/projects/Clinical-knowledge-embeddings.

Acknowledgements

R.J., B.Y.R., R.D.B., N.D., and M.Z. are supported by the Berkowitz Family Living Laboratory at Harvard Medical School and the Clalit Research Institute. R.J. and M.Z. gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, awards from Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH, Pfizer Research, Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders. We gratefully acknowledge the resources provided by the Clalit Research Institute. We thank Michelle M. Li and Ayush Noori for contributing to the model development and analyses.

Author contributions

R.J. developed a unified knowledge graph representation of clinical concepts and designed, implemented, and benchmarked heterogeneous graph transformer models. U.G. implemented the approach and conducted detailed analyses using CHS individual-level data. L.H., J.W., S.D., C.R.G., P.H., and R.S. formed an inter-institutional panel of clinicians that evaluated the alignment of clinical concept representations with established clinical knowledge across 90 diseases and 3,000 clinical codes. B.Y.R. and R.B. provided feedback on the methodology and evaluated the clinical concept embeddings through large-scale phenotype risk scoring and alignment with expert clinical knowledge. R.J., N.D., and M.Z. conceptualized and designed the study. All authors contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

References

1. National Research Council (US) Committee on A Framework for Developing a NewTaxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. (National Academies Press (US), Washington (DC), 2011).

2. Campbell, J. R. *et al.* Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity. *J. Am. Med. Inform. Assoc.* **4**, 238–251 (1997).

3. de Lusignan, S. Codes, classifications, terminologies and nomenclatures: definition, development and application in practice. *Inform. Prim. Care* **13**, 65–70 (2005).

4. Kather, J. N., Ferber, D., Wiest, I. C., Gilbert, S. & Truhn, D. Large language models could make natural language again the universal interface of healthcare. *Nat. Med.* 1–3 (2024) doi:10.1038/s41591-024-03199-w.

5. Hulsen, T. *et al.* From Big Data to Precision Medicine. *Front. Med.* **6**, 34 (2019).

6. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).

7. The Shaky Foundations of Foundation Models in Healthcare. https://hai.stanford.edu/news/shaky-foundations-foundation-models-healthcare (2023).

8. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* **15**, e1002683 (2018).

9. Yang, J., Soltan, A. A. S. & Clifton, D. A. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *Npj Digit. Med.* **5**, 1–8 (2022).

10. Panch, T. *et al.* “Yes, but will it work for my patients?” Driving clinically relevant research with benchmark datasets. *Npj Digit. Med.* **3**, 1–4 (2020).

11. Goetz, L., Seedat, N., Vandersluis, R. & van der Schaar, M. Generalization—a key challenge for responsible AI in patient-facing clinical applications. *Npj Digit. Med.* **7**, 1–4 (2024).

12. O’Malley, K. J. *et al.* Measuring Diagnoses: ICD Code Accuracy. *Health Serv. Res.* **40**, 1620 (2005).

13. Botsis, T., Hartvigsen, G., Chen, F. & Weng, C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit Transl. Bioinforma.* **2010**, 1–5 (2010).

14. Sudat, S. E., Robinson, S. C., Mudiganti, S., Mani, A. & Pressman, A. R. Mind the clinical-analytic gap: Electronic health records and COVID-19 pandemic response. *J. Biomed. Inform.* **116**, 103715 (2021).

15. Binkheder, S. *et al.* Real-World Evidence of COVID-19 Patients' Data Quality in the Electronic Health Records. *Healthcare* **9**, 1648 (2021).

16. Chalmers, R. J. G. Health Care Terminology for the Electronic Era. *Mayo Clin. Proc.* **81**, 729–731 (2006).

17. Song, Z. *et al.* Physician Practice Pattern Variations in Common Clinical Scenarios Within 5 US Metropolitan Areas. *JAMA Health Forum* **3**, e214698 (2022).

18. Baicker, K., Chandra, A., Skinner, J. S. & Wennberg, J. E. Who you are and where you live: how race and geography affect the treatment of medicare beneficiaries. *Health Aff. Proj. Hope* **Suppl Variation**, VAR33-44 (2004).

19. Baicker, K., Chandra, A. & Skinner, J. S. Geographic variation in health care and the problem of measuring racial disparities. *Perspect. Biol. Med.* **48**, S42-53 (2005).

20. Medicine, I. of, Services, B. on H. C. & Programs, C. on R. H. I. P. M., Payment, and Performance Improvement. *Rewarding Provider Performance: Aligning Incentives in Medicare*. (National Academies Press, 2007).

21. Blumenthal, D. & Tavenner, M. The "Meaningful Use" Regulation for Electronic Health Records. *N. Engl. J. Med.* **363**, 501–504 (2010).

22. McIlvennan, C. K., Eapen, Z. J. & Allen, L. A. Hospital Readmissions Reduction Program. *Circulation* **131**, 1796–1803 (2015).

23. Morden, N. E., Colla, C. H., Sequist, T. D. & Rosenthal, M. B. Choosing Wisely — The Politics and Economics of Labeling Low-Value Services. *N. Engl. J. Med.* **370**, 589–592 (2014).

24. Fisher, E. S. *et al.* The implications of regional variations in Medicare spending. Part 1: the content, quality, and accessibility of care. *Ann. Intern. Med.* **138**, 273–287 (2003).

25. Charlesworth, C. J., Meath, T. H. A., Schwartz, A. L. & McConnell, K. J. Comparison of Low-Value Care in Medicaid vs Commercially Insured Populations. *JAMA Intern. Med.* **176**, 998–1004 (2016).

26. Baicker, K., Chernew, M. E. & Robbins, J. A. The spillover effects of Medicare managed care: Medicare Advantage and hospital utilization. *J. Health Econ.* **32**, 1289–1300 (2013).

27. Ancker, J. S. *et al.* How is the electronic health record being used? Use of EHR data to assess physician-level variability in technology use. *J. Am. Med. Inform. Assoc.* **21**, 1001–1008 (2014).

28. Roland, M. Linking Physicians' Pay to the Quality of Care — A Major Experiment in the United Kingdom. *N. Engl. J. Med.* **351**, 1448–1454 (2004).

29. Sirovich, B. E., Gottlieb, D. J., Welch, H. G. & Fisher, E. S. Variation in the Tendency of Primary Care Physicians to Intervene. *Arch. Intern. Med.* **165**, 2252–2256 (2005).

30. Berner, E. S., Kasiraman, R. K., Yu, F., Ray, M. N. & Houston, T. K. Data quality in the outpatient setting: impact on clinical decision support systems. *AMIA Annu. Symp. Proc. AMIA Symp.* 41–45 (2005).

31. Verheij, R. A., Curcin, V., Delaney, B. C. & McGilchrist, M. M. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J. Med. Internet Res.* **20**, e185 (2018).

32. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479 (2018).

33. Tang, P. C., Ralston, M., Arrigotti, M. F., Qureshi, L. & Graham, J. Comparison of Methodologies for Calculating Quality Measures Based on Administrative Data versus Clinical Data from an Electronic Health Record System: Implications for Performance Measures. *J. Am. Med. Inform. Assoc.* **14**, 10–15 (2007).

34. Hogan, W. R. & Wagner, M. M. Accuracy of Data in Computer-based Patient Records. *J. Am. Med. Inform. Assoc.* **4**, 342 (1997).

35. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc. JAMIA* **20**, 117 (2012).

36. Williams, R., Kontopantelis, E., Buchan, I. & Peek, N. Clinical code set engineering for reusing EHR data for research: A review. *J. Biomed. Inform.* **70**, 1–13 (2017).

37. Delvaux, N. *et al.* Coding Systems for Clinical Decision Support: Theoretical and Real-World Comparative Analysis. *JMIR Form. Res.* **4**, e16094 (2020).

38. Odigie, E. *et al.* Fast Healthcare Interoperability Resources, Clinical Quality Language, and Systematized Nomenclature of Medicine—Clinical Terms in Representing Clinical Evidence Logic Statements for the Use of Imaging Procedures: Descriptive Study. *JMIR Med. Inform.* **7**, e13590 (2019).

39. Kokkinakis, D. What is the Coverage of SNOMED CT&#174;on Scientific Medical Corpora? in *User Centred Networked Health Care* 814–818 (IOS Press, 2011). doi:10.3233/978-1-60750-806-9-814.

40. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).

41. Yang, J. *et al.* Poisoning medical knowledge using large language models. *Nat. Mach. Intell.* **6**, 1156–1168 (2024).

42. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *Npj Digit. Med.* **6**, 1–4 (2023).

43. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).

44. Soroush, A. *et al.* Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. *NEJM AI* **1**, AIdbp2300040 (2024).

45. Huang, J. *et al.* A critical assessment of using ChatGPT for extracting structured data from clinical notes. *Npj Digit. Med.* **7**, 1–13 (2024).

46. Burford, K. G., Itzkowitz, N. G., Ortega, A. G., Teitler, J. O. & Rundle, A. G. Use of Generative AI to Identify Helmet Status Among Patients With Micromobility-Related Injuries From Unstructured Clinical Notes. *JAMA Netw. Open* **7**, e2425981 (2024).

47. Lee, S. A. & Lindsey, T. Can Large Language Models abstract Medical Coded Language? Preprint at https://doi.org/10.48550/arXiv.2403.10822 (2024).

48. The CORE Problem List Subset of SNOMED CT®. https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html.

49. von Rueden, L. *et al.* Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Trans. Knowl. Data Eng.* **35**, 614–633 (2023).

50. Sirocchi, C., Bogliolo, A. & Montagna, S. Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Med. Inform. Decis. Mak.* **24**, 186 (2024).

51. General Cardiovascular Risk Profile for Use in Primary Care | Circulation. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.107.699579.

52. ACMG recommendations for standards for interpretation of sequence variations. *Genet. Med.* **2**, 302–303 (2000).

53. Poonacha, T. K. & Go, R. S. Scientific evidence underlying National Comprehensive Cancer Network Clinical Practice Guidelines. *J. Clin. Oncol.* **28**, 6020–6020 (2010).

54. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. *Clinical Practice Guidelines We Can Trust*. (National Academies Press (US), Washington (DC), 2011).

55. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).

56. Nelson, W. *et al.* To Embed or Not: Network Embedding as a Paradigm in Computational Biology. *Front. Genet.* **10**, (2019).

57. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).

58. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf. Fusion* **50**, 71–91 (2019).

59. Zhou, H. *et al.* Knowledge-guided convolutional networks for chemical-disease relation extraction. *BMC Bioinformatics* **20**, 260 (2019).

60. Nunes, S., Sousa, R. T. & Pesquita, C. Multi-domain knowledge graph embeddings for gene-disease association prediction. *J. Biomed. Semant.* **14**, 11 (2023).

61. Himmelstein, D. S. & Baranzini, S. E. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLOS Comput. Biol.* **11**, e1004259 (2015).

62. Morris, J. H. *et al.* The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* **39**, btad080 (2023).

63. Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol.* **40**, 692–702 (2022).

64. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, e26726 (2017).

65. Gogleva, A. *et al.* Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat. Commun.* **13**, 1667 (2022).

66. Zhu, Q. *et al.* An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). *J. Biomed. Semant.* **11**, 13 (2020).

67. Sosa, D. N. *et al.* A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **25**, 463–474 (2020).

68. Feng, F. *et al.* GenomicKB: a knowledge graph for the human genome. *Nucleic Acids Res.* **51**, D950–D956 (2023).

69. Renaux, A. *et al.* A knowledge graph approach to predict and interpret disease-causing gene interactions. *BMC Bioinformatics* **24**, 324 (2023).

70. Percha, B. & Altman, R. B. A global network of biomedical relationships derived from text. *Bioinformatics* **34**, 2614–2624 (2018).

71. Ernst, P., Siu, A. & Weikum, G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics* **16**, 157 (2015).

72. Zheng, S. *et al.* PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Brief. Bioinform.* **22**, bbaa344 (2021).

73. Robinson, P. N. *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).

74. Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world. 2022.04.13.22273750 Preprint at https://doi.org/10.1101/2022.04.13.22273750 (2022).

75. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci. Data* **10**, 67 (2023).

76. Wang, M. *et al.* PDD Graph: Bridging Electronic Medical Records and Biomedical Knowledge Graphs via Entity Linking. in *The Semantic Web – ISWC 2017* (eds. d'Amato, C. et al.) 219–227 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-68204-4_23.

77. Nelson, C. A., Butte, A. J. & Baranzini, S. E. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat. Commun.* **10**, 3045 (2019).

78. Huan, J.-M. *et al.* The biomedical knowledge graph of symptom phenotype in coronary artery plaque: machine learning-based analysis of real-world clinical data. *BioData Min.* **17**, 13 (2024).

79. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S. & Sontag, D. Learning a Health Knowledge Graph from Electronic Medical Records. *Sci. Rep.* **7**, 5994 (2017).

80. Murali, L., Gopakumar, G., Viswanathan, D. M. & Nedungadi, P. Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study. *J. Biomed. Inform.* **143**, 104403 (2023).

81. Zheng, N. S. *et al.* PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. *J. Am. Med. Inform. Assoc.* **27**, 1675–1687 (2020).

82. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).

83. ATCDDD - Structure and principles. https://atcddd.fhi.no/atc/structure_and_principles/.

84. RxNorm. https://www.nlm.nih.gov/research/umls/rxnorm/index.html.

85. Hu, Z., Dong, Y., Wang, K. & Sun, Y. Heterogeneous Graph Transformer. Preprint at https://doi.org/10.48550/arXiv.2003.01332 (2020).

86. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).

87. Xie, Y., Xu, Z., Zhang, J., Wang, Z. & Ji, S. Self-Supervised Learning of Graph Neural Networks: A Unified Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 2412–2429 (2023).

88. Oord, A. van den, Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. Preprint at https://doi.org/10.48550/arXiv.1807.03748 (2019).

89. Kahaly, G. J. *et al.* 2018 European Thyroid Association Guideline for the Management of Graves' Hyperthyroidism. *Eur. Thyroid J.* **7**, 167–186 (2018).

90. Lab Tests for Lupus | Lupus Foundation of America. https://www.lupus.org/resources/lab-tests-for-lupus.

91. Anderson, I. M. & Edwards, J. G. Guidelines for choice of selective serotonin reuptake inhibitor in depressive illness. *Adv. Psychiatr. Treat.* **7**, 170–180 (2001).

92. Kirby, J. C. *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc. JAMIA* **23**, 1046–1052 (2016).

93. Mikolov, T., Yih, S. W. & Zweig, G. Linguistic Regularities in Continuous Space Word Representations. in (2013).

94. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

95. Conrad, N. *et al.* Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK. *Lancet Lond. Engl.* **401**, 1878–1890 (2023).

96. Heart disease - Symptoms and causes. *Mayo Clinic* https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

97. COPD - Symptoms and causes. *Mayo Clinic* https://www.mayoclinic.org/diseases-conditions/copd/symptoms-causes/syc-20353679.

98. Alzheimer's disease - Symptoms and causes. *Mayo Clinic* https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447.

99. Lung cancer - Symptoms and causes. *Mayo Clinic* https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620.

100. Stroke - Symptoms and causes. *Mayo Clinic* https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113.

101. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genet.* **9**, e1003348 (2013).

102. Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233–1239 (2018).

103. Balicer, R. D. & Afek, A. Digital health nation: Israel's global big data innovation hub. *The Lancet* **389**, 2451–2453 (2017).

104. Rennert, G. & Peterburg, Y. Prevalence of selected chronic diseases in Israel. *Isr. Med. Assoc. J. IMAJ* **3**, 404–408 (2001).

105. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. A guide to performing Polygenic Risk Score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

106. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html.

107. Murphy, D. L. *et al.* Anxiety and affective disorder comorbidity related to serotonin and other neurotransmitter systems: obsessive–compulsive disorder as an example of overlapping clinical and genetic heterogeneity. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120435 (2013).

108. Dash, T., Chitlangia, S., Ahuja, A. & Srinivasan, A. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Sci. Rep.* **12**, 1040 (2022).

109. Sheth, A., Gaur, M., Kursuncu, U. & Wickramarachchi, R. Shades of Knowledge-Infused Learning for Enhancing Deep Learning. *IEEE Internet Comput.* **23**, 54–63 (2019).

110. Leiser, F., Rank, S., Schmidt-Kraepelin, M., Thiebes, S. & Sunyaev, A. Medical informed machine learning: A scoping review and future research directions. *Artif. Intell. Med.* **145**, 102676 (2023).

111. Choi, E., Xiao, C., Stewart, W. F. & Sun, J. MiME: multilevel medical embedding of electronic health records for predictive healthcare. in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* 4552–4562 (Curran Associates Inc., Red Hook, NY, USA, 2018).

112. Choi, E., Bahadori, M. T., Song, L., Stewart, W. F. & Sun, J. GRAM: Graph-based Attention Model for Healthcare Representation Learning. Preprint at https://doi.org/10.48550/arXiv.1611.07012 (2017).

113. Ma, F. *et al.* KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare. in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* 743–752 (Association for Computing Machinery, New York, NY, USA, 2018). doi:10.1145/3269206.3271701.

114. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Npj Digit. Med.* **4**, 1–13 (2021).

115. Finch, A. *et al.* Exploiting hierarchy in medical concept embedding*. *JAMIA Open* **4**, ooab022 (2021).

116. Nelson, C. A., Bove, R., Butte, A. J. & Baranzini, S. E. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *J. Am. Med. Inform. Assoc.* **29**, 424–434 (2022).

117. Steiger, E. & Kroll, L. E. Patient Embeddings From Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework. *JMIR AI* **2**, e40755 (2023).

118. Matta, A. *et al.* Embedding Representations of Diagnosis Codes for Outlier Payment Detection. in *2023 International Conference on Machine Learning and Applications (ICMLA)* 2153–2160 (2023). doi:10.1109/ICMLA58977.2023.00325.

119. Wu, T., Wang, Y., Wang, Y., Zhao, E. & Yuan, Y. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. *Sci. Rep.* **11**, 5858 (2021).

120. Steinberg, E. *et al.* Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inform.* **113**, 103637 (2021).

121. Choi, Y., Chiu, C. Y.-I. & Sontag, D. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Summits Transl. Sci. Proc.* **2016**, 41 (2016).

122. Zou, Y. *et al.* Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. *Sci. Rep.* **12**, 17868 (2022).

123. Jiang, P., Xiao, C., Cross, A. & Sun, J. GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs. Preprint at https://doi.org/10.48550/arXiv.2305.12788 (2024).

124. Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *Npj Digit. Med.* **3**, 1–11 (2020).

125. Hamilton, W. L., Ying, R. & Leskovec, J. Representation Learning on Graphs: Methods and Applications. Preprint at https://doi.org/10.48550/arXiv.1709.05584 (2018).

126. Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022).

127. OpenAI *et al.* GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).

128. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).

129. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Preprint at https://doi.org/10.48550/arXiv.2005.11401 (2021).

130. Lukas, N. *et al.* Analyzing Leakage of Personally Identifiable Information in Language Models. in *2023 IEEE Symposium on Security and Privacy (SP)* 346–363 (2023). doi:10.1109/SP46215.2023.10179300.

131. Pan, X., Zhang, M., Ji, S. & Yang, M. Privacy Risks of General-Purpose Language Models. in *2020 IEEE Symposium on Security and Privacy (SP)* 1314–1331 (2020). doi:10.1109/SP40000.2020.00095.

132. Heart Disease in Men and Women. https://www.massgeneralbrigham.org/en/about/newsroom/articles/heart-disease-in-men-and-women.

133. Shuey, M. M. *et al.* Next-generation phenotyping: introducing phecodeX for enhanced discovery research in medical phenomics. *Bioinformatics* **39**, btad655 (2023).

134. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789-798 (2015).

135. Shevtsova, D. *et al.* Trust in and Acceptance of Artificial Intelligence Applications in Medicine: Mixed Methods Study. *JMIR Hum. Factors* **11**, e47031 (2024).

136. International Classification of Diseases (ICD). https://www.who.int/standards/classifications/classification-of-diseases.

137. CPT® overview and code approval. *American Medical Association* https://www.ama-assn.org/practice-management/cpt/cpt-overview-and-code-approval (2024).

138. Forrey, A. W. *et al.* Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin. Chem.* **42**, 81–90 (1996).

139. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* **26**, 1205–1210 (2010).

140. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* **12**, e0175508 (2017).

141. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).

142. Knowledge Base. *LOINC* https://loinc.org/kb/.

143. McInnes, B. T., Pedersen, T. & Pakhomov, S. V. S. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. *AMIA. Annu. Symp. Proc.* **2009**, 431–435 (2009).

144. Wilcke, X., Bloem, P. & de Boer, V. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Sci.* **1**, 39–57 (2017).

145. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. Preprint at https://doi.org/10.48550/arXiv.2103.00020 (2021).

146. Sohn, K. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. in *Advances in Neural Information Processing Systems* vol. 29 (Curran Associates, Inc., 2016).

147. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. in *Proceedings of the 37th International Conference on Machine Learning* vol. 119 1597–1607 (JMLR.org, 2020).

148. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2018).

149. PheWAS - Phenome Wide Association Studies. https://phewascatalog.org/phecodes_v1_1.

150. Anxiety algorithm | PheKB. https://phekb.org/phenotype/anxiety-algorithm.

151. Chronic Kidney Disease | PheKB. https://phekb.org/phenotype/chronic-kidney-disease.

152. Breast Cancer | PheKB. https://phekb.org/phenotype/breast-cancer.

153. Medical Diseases & Conditions. *Mayo Clinic* https://www.mayoclinic.org/diseases-conditions.

154. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **22** **140**, 55–55 (1932).