ClinVec: Unified Embeddings of Clinical Codes Enable Knowledge-Grounded AI in 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], Xiaorui Su[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 structured clinical knowledge into artificial intelligence (AI) models remains a major challenge. Medical codes primarily reflect administrative workflows rather than clinical reasoning, limiting AI models' ability to capture true clinical relationships and undermining their generalizability. To address this, we introduce *ClinGraph*, a clinical knowledge graph that integrates eight EHR-based vocabularies, and *ClinVec*, a set of 153,166 clinical code embeddings derived from *ClinGraph* using a graph transformer neural network. *ClinVec* provides a machine-readable representation of clinical knowledge that captures semantic relationships among diagnoses, medications, laboratory tests, and procedures. Panels of clinicians from multiple institutions evaluated the embeddings across 96 diseases and more than 3,000 clinical codes, confirming their alignment with expert knowledge. In a retrospective analysis of 4.57 million patients from Clalit Health Services, we show that *ClinVec* supports phenotype risk scoring and stratifies individuals by survival outcomes. We further demonstrate that injecting *ClinVec* into large language models improves performance on medical question answering, including for region-specific clinical scenarios. *ClinVec* enables structured clinical knowledge to be injected into predictive and generative AI models, bridging the gap between EHR codes and clinical reasoning.

Introduction

Medicine builds on centuries of accumulated knowledge and the goal of individualized patient care through clinical reasoning and evidence-based practice[1]. Over time, medical vocabularies and ontologies have been developed to represent clinical information, fostering interoperability and data exchange across healthcare systems[2–4]. However, these standardized coding systems remain fragmented, with each vocabulary designed for specific use cases. This fragmentation limits their integration into artificial intelligence (AI) models and introduces a gap between clinical knowledge and its computational representation in precision medicine.

The widespread adoption of electronic health records (EHRs) and data standardization initiatives has accelerated the use of structured clinical data in AI models[5,6], with more than half of healthcare foundation AI models relying on structured clinical codes, such as billing data and medication records[7]. These models implicitly assume that EHRs contain information sufficient to capture the clinical concepts relevant to patient care. In practice, this assumption often fails. Models trained on EHR data frequently learn patterns specific to a single healthcare system rather than generalizable medical relationships[8–11]. The variability of clinical codes across institutions further limits the transferability of prediction models[12–15]. Many codes are selected or recorded for operational rather than clinical reasons[16], influenced by geographic factors[17–19], national policies[20–23], insurance requirements[24–26], physician practices[27–29], and other features of local healthcare delivery[30–34]. EHR-based models, as a result, often fail to reflect consistent clinical definitions or relationships among codes[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]. LLMs can struggle to 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 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.

Most AI models lack mechanisms to integrate clinical knowledge, limiting their ability to align with medical best practices[49,50]. Clinical researchers have published over 3,700 evidence-based guidelines across 39 countries[51–54], yet this structured knowledge remains largely unused in model development. Most AI models rely on statistical patterns in patient data or text corpora, but these approaches do not capture the clinical knowledge embedded in structured EHR codes. Bridging this gap requires consistent, machine-readable representations that reflect how clinical codes relate to real-world medical practice. While embeddings exist for text and images[55–60], no comparable resource captures structured clinical knowledge for use in AI models.

To address the gap between clinical knowledge and structured EHR data, we developed a *ClinGraph*, a clinical knowledge graph spanning eight standardized medical vocabularies, and *ClinVec*, a set of 153,166 clinical code embeddings derived from *ClinGraph*. *ClinVec* defines a unified, machine-readable latent space that captures relationships among diagnoses, medications, laboratory tests, and procedures. This approach generates clinically meaningful representations of medical codes in a hypothesis-free manner and provides a resource for integrating structured clinical knowledge into predictive models for precision medicine. The learned embeddings reveal patterns aligned with human anatomy and disease phenotypes and were validated by panels of clinical experts across 96 diseases and more than 3,000 clinical codes. We further demonstrate the predictive utility of *ClinVec* through phenotype risk score analyses for chronic kidney disease, chronic obstructive pulmonary disease, and heart failure, using EHR data from 4.57 million patients in Clalit Health Services. Finally, we show that integrating *ClinVec* into large language models improves medical question answering across a range of clinical scenarios. *ClinGraph* and *ClinVec* provide a foundation for embedding structured clinical knowledge into AI models, supporting interpretability, cross-institutional transferability, and alignment with clinical practice.

Results

ClinGraph and *ClinVec*: Representing structured medical knowledge for AI

Many efforts have focused on structuring clinical knowledge for enhanced interoperability and data harmonization. Clinical knowledge has been traditionally structured using Common Data Models (CDMs), which standardize data by mapping diverse coding systems into a unified schema[61]. Alternatively, biomedical ontologies organize information in a hierarchical structure to facilitate semantic interoperability and reasoning across coding systems. However, both rely on fixed schemas and expert-defined relationships that limit their contextual richness and

adaptability to real-world data. KGs offer a more flexible approach by integrating diverse knowledge sources without rigid schema constraints. However, most knowledge graphs do not explicitly incorporate standard clinical vocabularies and instead focus on biological processes[62–64], therapeutics[55,65,66], genetic phenotypes[67–70], or a combination thereof[71–73]. KGs that incorporate clinical vocabularies are often restricted to ICD codes (diagnosis codes) or rely on medical terminologies like HPO[74] and MONDO[75], which are not standard in most EHRs[76–78]. Several approaches construct KGs directly from patient data[79–81], but this process is highly sensitive to the source patient population, potentially limiting generalizability.

We develop a resource consisting of *ClinGraph*, a knowledge graph consisting of 153,166 nodes and 2,873,025 edges spanning eight clinical vocabularies (Figure 1A, 1B), and *ClinVec*, a set of high-dimensional, vector representations (embeddings) of clinical knowledge that can be integrated into AI models (Figure 1C, 1D). *ClinGraph* incorporates diagnosis codes (ICD9CM, ICD10CM, phecodes), laboratory tests (LOINC), medications (RxNorm, ATC), procedures (CPT), and other controlled clinical vocabularies (SNOMED CT, UMLS), and represents each concept in the KG as a medical code (Supplementary Table S1). Relational information is derived from multiple databases, including ontologies such as Unified Medical Language System[82] (UMLS) and empirically constructed knowledge bases such as PheMap[83], which derives relationships from the large-scale mining of biomedical literature (Supplementary Table S2).

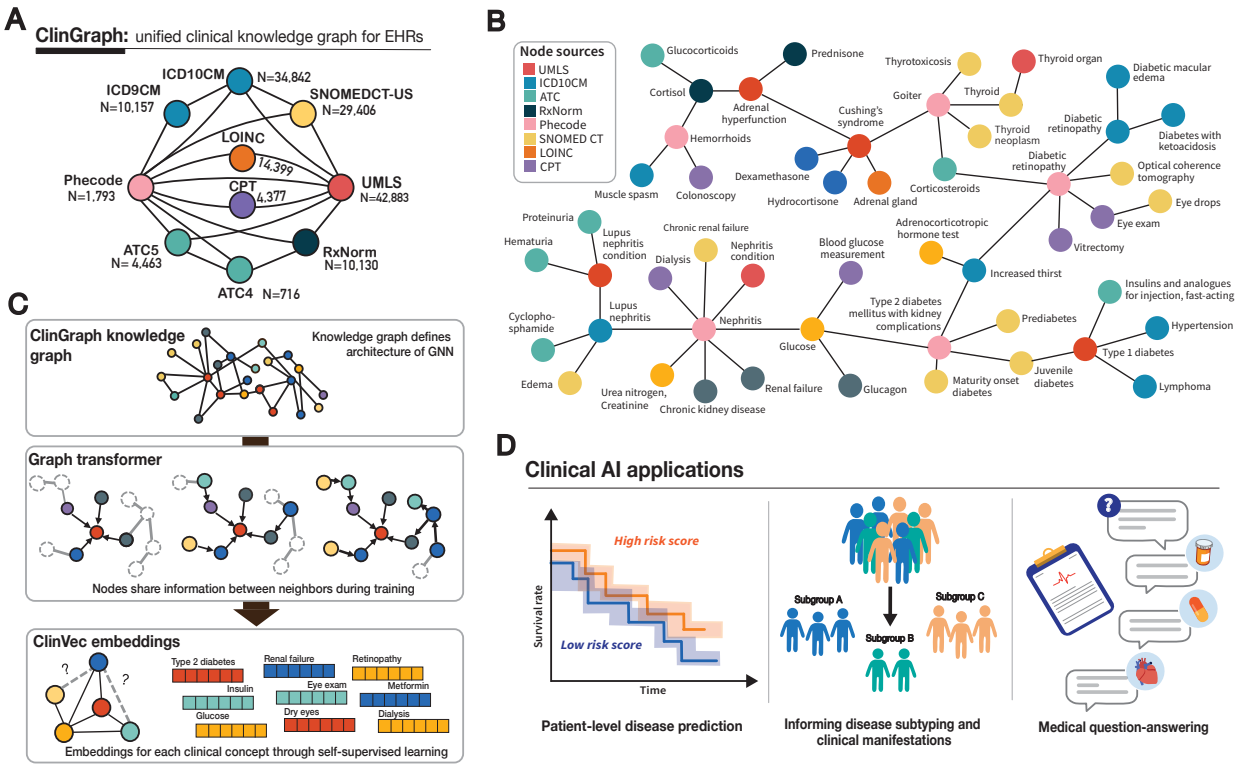

Figure 1: Overview of approach. (A) Meta-graph showing the structure of *ClinGraph* and the edge relationships between clinical vocabularies shown as node types. **(B)** An example local neighborhood of *ClinGraph*. Each clinical code is represented as a node, and the color denotes the clinical vocabulary. **(C)** Workflow for generating clinical knowledge embeddings: building the clinical knowledge graph, constructing a graph transformer model, and generating embeddings through self-supervised learning. **(D)** Clinical AI applications of *ClinGraph* and *ClinVec*.

A key feature of *ClinGraph* is that it represents clinical concepts using their original coding systems, rather than aggregating them into a common vocabulary. This design preserves coding-specific structures such as hierarchical relationships and clinically meaningful groupings and avoids the information loss that can occur when mapping between vocabularies. To evaluate the coverage of EHR coding systems, we compared *ClinGraph* to six existing KGs. For each KG, we mapped its node identifiers to EHR codes and calculated the proportion of coverage for common coding systems (Supplementary Table S3). *ClinGraph* is the only resource that integrates diagnoses, medications, laboratory tests, and procedures in a unified framework based on structured EHR data. Other KGs do not include nodes for procedures or laboratory tests. While *ClinGraph*, PrimeKG[76], and PharMeBINet[84] capture similar proportions of ICD10CM codes, *ClinGraph* spans a broader range of diagnoses. By evaluating the number of distinct ICD10CM root codes (for example, J45), we found that *ClinGraph* includes 35% more disease areas than other graphs. This likely reflects the clinical orientation of *ClinGraph*, in contrast to other resources that emphasize biological and genetic phenotypes.

While KGs represent information as discrete nodes and edges, embedding-based representations create a machine-readable form of information that encodes clinical concepts and their contextual information into continuous vectors. Transforming discrete KGs into continuous embeddings enables complex inference tasks, such as grounding AI models in clinical knowledge[85], that are not possible using only node-and-edge graph representations. Using a heterogenous graph transformer (HGT)[86], we learn a function that maps each node (medical code) in the KG to low-dimensional representations (embeddings). The KG determines how information is shared between nodes throughout the KG, while a multi-head attention mechanism enhances the model's ability to up-weight and down-weight different parts of the topological space[87]. We model four distinct node types—diagnosis codes, medications, laboratory tests/procedures, and clinical concepts/vocabularies—where grouping similar vocabularies (e.g., RxNorm and ATC) enables information sharing across related nodes while preserving enough type-specific information to support effective parameter learning (Methods). We train the model for edge prediction through self-supervised learning via contrastive edge masking[88]. This approach generates embeddings for each clinical code such that codes located near one another in the knowledge graph are also positioned closely in the latent space, reflecting their semantic similarity.

To support the integration of *ClinVec* into AI models, the learned embeddings must be consistent with current clinical knowledge and reflect information used in routine healthcare. Existing reference datasets for evaluating the semantic relatedness of clinical terms are limited in size (fewer than 1,500 data points) and based on outdated coding systems that include obsolete codes. We constructed a new validation dataset consisting of 96 diseases across 10 major disease categories, such as circulatory and respiratory disorders, and over 3,000 SNOMED CT codes. A panel of expert clinicians from Israel and the United States annotated the dataset. Based on clinician input, we selected 96 phecodes representing conditions that are clinically relevant to general adult patient populations. For each phecode, we identified the 20 SNOMED CT codes with the highest cosine similarity to the phecode embedding and randomly sampled an additional 20 SNOMED CT codes as controls (Figure 2A).

Codes already forming an edge with the disease node within the knowledge graph were excluded from the analysis. The final evaluation dataset comprises 96 disease lists, each with 40 SNOMED CT codes for a total of 3,680 data points, almost 3 times the size of existing benchmarks combined.

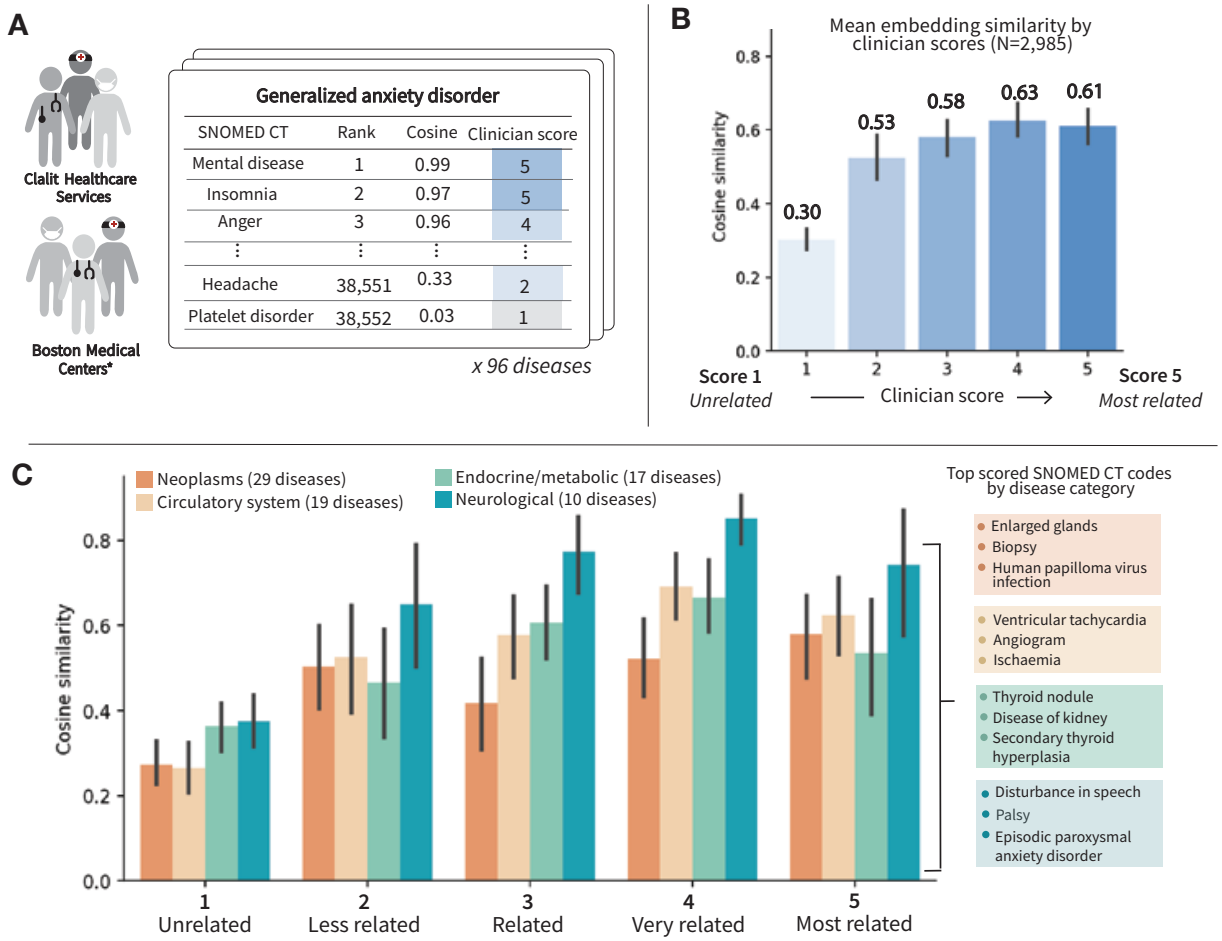

Figure 2: Concordance between *ClinVec* embedding similarity and clinical expert scores. (A) Overview of clinical expert review process. **(B)** We report the mean cosine similarity across all disease-term pairs stratified by clinical expert scores 1-5. **(C)** We stratify cosine similarity scores across four disease categories and show examples of SNOMED CT codes that were assigned a clinical expert score of 5.

To assess the relevance of each clinical code to a given disease, clinicians graded SNOMED CT codes on a scale from 1 (unrelated) to 5 (very relevant), evaluating both accuracy and applicability to clinical practice (Figure 2B). Clinicians graded each SNOMED CT code on a scale from 1 to 5, where 1 indicated the code was unrelated to the target disease and 5 indicated it was highly relevant and frequently observed in clinical practice. We compared these expert ratings to the cosine similarity between the corresponding *ClinVec* embeddings. Grouping similarity scores by clinician rating revealed a positive correlation, with mean cosine similarity increasing across scores 1 to 5 (0.30, 0.53, 0.58, 0.63, 0.61). When stratifying by disease category, we observed lower alignment for neoplasms compared to other categories. In this case, high embedding similarity was often assigned a low clinical relevance score by reviewers. This misalignment likely reflects the heterogeneity of neoplasms, which vary by organ system, stage, and metastatic

behavior. Among disease-code pairs with low clinician scores (≤2) and high cosine similarity (>0.8), many mismatches involved anatomical proximity rather than clinical relevance. For example, top-scoring terms for "cancer of salivary glands" included thyroid conditions. Although these organs are anatomically adjacent and involved in autoimmune diseases such as Sjögren's syndrome or Hashimoto's thyroiditis, they are not directly related in the context of salivary gland cancer. These findings suggest that embeddings for highly specific diseases could benefit from fine-tuning on subgraphs of the KG to improve granularity and clinical alignment.

Anatomical and medical subspecialty structure emerge in the embedding space

Clinical knowledge embeddings must capture patterns that align with human biology, disease presentation, and clinical workflows. Although *ClinGraph* shares some content with other biomedical KGs, its integration of clinical vocabularies and relationship types produces embeddings that more effectively represent clinically meaningful structure in the latent space. To evaluate this, we trained heterogeneous graph transformers on PrimeKG and a SNOMED CT-based KG using the same model architecture and training setup as for *ClinGraph*, and compared the resulting embedding spaces (Methods).

We characterize the breadth and granularity of the clinical knowledge represented in the learned embeddings. Visualizing the latent space of disease (Figure 3A, Supplementary Figure S1A-C) and drug embeddings (Figure 3B, Supplementary Figure S1D-F) generated by each KG, the embeddings derived from *ClinGraph* broadly maintain groupings of diseases by organ system, mirroring the organization of clinical care. However, embeddings produced by PrimeKG and SNOMED-CT KG do not exhibit this distinct clustering pattern. This observation is supported when comparing silhouette scores and adjusted mutual information (AMI) where embedding clusters from *ClinGraph* (silhouette: 0.54, AMI: 0.31) produce metrics at the same value or greater than clusters produced from other KGs (PrimeKG - silhouette: 0.54, AMI: 0.02; SNOMED-CT KG - silhouette: 0.15, AMI: -1e-4). Further examining the clusters of phecode embeddings from *ClinGraph*, although some phenotypes deviate from these main clusters, many deviations reflect known distinctions within clinical care. For example, most endocrine conditions cluster together (Figure 3A: orange cluster), but some phenotypes cluster more closely with sense organ conditions (Figure 3A: 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.

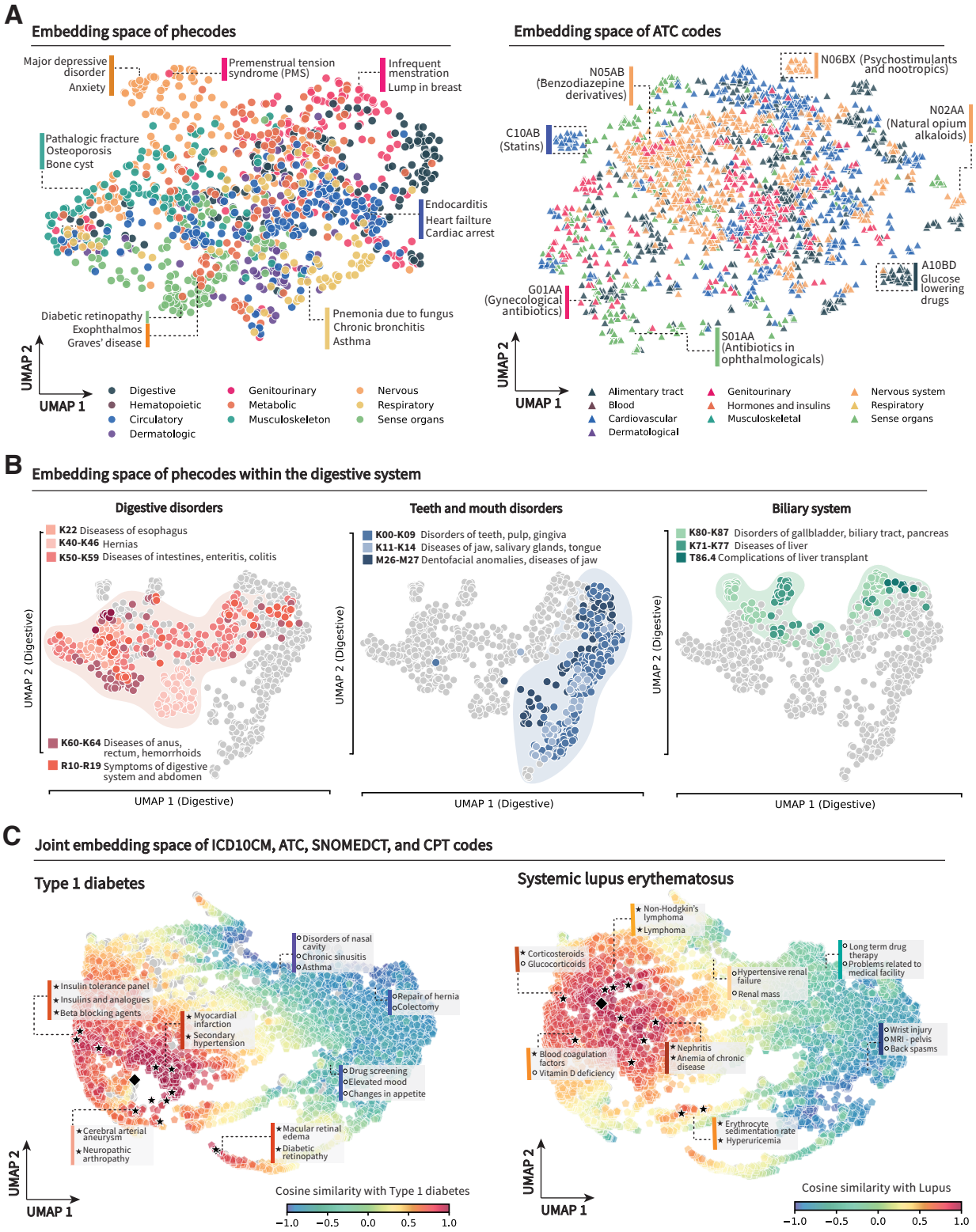

Figure 3. Latent *ClinVec* space captures biomedical knowledge consistent with organ system biology and anatomy. (A) Scatter plots showing a reduced dimension of the embedding space of clinical codes. On the left, the latent space of phecodes is shown, where each phecode is shaded by its corresponding organ system categorization. On the right, the latent space of ATC medication codes is shown, where each ATC code is shaded according to the corresponding ATC level 1. **(B)** Visualization of the embedding region of digestive system ICD10CM codes. We

highlight the following phenotype subcategories: functional digestive disorders – phecodes, teeth and mouth disorders (center), and biliary system (right). **(C)** Visualization of embedding space overlaid with ICD10CM, ATC, SNOMEDCT, and CPT codes. Codes are shaded based on cosine similarity with type 1 diabetes (left) and systematic lupus erythematosus (right), and the target disease embedding is denoted with a black diamond. The black stars indicate a subset of codes related to each disease. These are presented along with other codes in their proximity but are not directly related to the disease of interest.

Extending this clustering analysis to drug categories based on the ATC-1 level, embeddings generated by both *ClinGraph* and PrimeKG show clustering by drug category although *ClinGraph* has a higher silhouette score (0.71 vs. 0.64). In contrast, embeddings from the SNOMED-CT knowledge graph do not produce distinct clusters and have an AMI score close to zero. This is likely attributable to the more extensive drug coverage in *ClinGraph* compared to SNOMED-CT KG. Again, we also find that deviations in the expected cluster labeling according to the ATC-1 classification in *ClinGraph* 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. These analyses demonstrate that the effectiveness of clinical embeddings depends on the design of the underlying KG. *ClinGraph*'s integration of clinically grounded vocabularies and relationship types results in more accurate representations, outperforming existing KGs across multiple evaluation metrics.

A 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 *ClinVec* clinical vocabulary embeddings, we perform a separate dimensionality reduction analysis using only phecodes of the digestive system (Figure 3B). 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 ($p\text{-}value=3.65\times10^{-45}$; two-sided Mann-Whitney U test), biliary system ($p\text{-}value=6.38\times10^{-44}$), and conditions of mouth and teeth ($p\text{-}value=1.05\times10^{-74}$) 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 show that *ClinVec* embeddings to capture fine-grained distinctions within broader medical categories.

A key advantage of *ClinVec* embedding space is that it allows direct comparison of clinical codes across different vocabularies, bridging gaps between otherwise disconnected coding

256 systems. We compare *ClinVec* to two state-of-the-art open-source clinical embeddings,
257 snomed2vec[89] and cui2vec[90], in a set of zero-shot retrieval tasks. For a given disease, we select
258 the top k closest codes within the embedding space and compare with these with the diseases'
259 symptom list as listed in the Mayo Clinic disease description. We report each model's perfor-
260 mance using 'hits @ k' which measures the proportion of diseases where a relevant symptom is
261 ranked among the top k items within the embedding space. We find that the *ClinVec* embeddings
262 outperform other approaches across a variety of values of *k* (Supplementary Table S4A). While
263 hierarchical structures explain some of the clustering, the embeddings also capture meaningful
264 patterns in cross-domain relationships, such as between diseases and their associated treatments,
265 which are not strictly hierarchical. We perform a similar analysis for drug indications with estab-
266 lished first-line drug treatments. For example, non-Hodgkin's lymphoma and CHOP (cyclophos-
267 phamide, doxorubicin, vincristine, prednisone), testicular cancer and BEP (bleomycin, etoposide,
268 cisplatin), and major depressive disorder and selective serotonin reuptake inhibitors. Limiting re-
269 trieval over codes exclusively describing ATC5-level medications, for each disease, we perform
270 zero-shot retrieval across >2,000 drug codes and assess the codes with the highest cosine similar-
271 ity with the disease embedding. We find that the *ClinVec* embeddings outperform other ap-
272 proaches across a variety of values of *k* (Supplementary Table S4B).

273 **Embedding space reveals clinically grounded diagnostic and symptom patterns**

274 Many clinical applications require disease-level resolution, where models must distinguish be-
275 tween specific diagnoses and their associated symptom profiles. To evaluate how *ClinVec* cap-
276 tures these diagnostic patterns, we examine type 1 diabetes (T1D) and systemic lupus erythema-
277 tosus (Figure 3C). We compute the cosine similarity between the T1D embedding and ICD10CM
278 codes, CPT, LOINC, and SNOMED CT codes within the embedding space and visualize these
279 patterns by projecting the computed cosine similarity onto the latent space. Broadly, near the em-
280 bedding of T1D, we observe regions of high cosine similarity as shaded in red and orange. We
281 find that many codes within this region are consistent with relevant T1D laboratory tests, such as
282 'insulin tolerance panel', and common complications, including cardiovascular diseases (Meth-
283 ods). We also identify a cluster of eye-related tests for T1D, including codes for eye exams and
284 macular edema. This highlights the embedding space's ability to capture nuanced patterns that
285 extend beyond identifying codes associated with the same organ system, recapitulating patterns
286 aligned with established disease presentations. Repeating this process for lupus, regions with

high cosine similarity highlight relevant tests such as erythrocyte sedimentation rate, which is used to assess systemic inflammation and nephritis, a common complication.

ClinVec contains the expressivity and flexibility to describe a wide range of clinical scenarios within the embedding space. Through embedding arithmetic, embeddings can be combined mathematically to form new meanings. For 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. These observations mirror the geometric properties of word embeddings in natural language processing[91], effectively capturing the relationship between concepts without explicitly defining it in the original vocabulary. 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 4A, 4B).

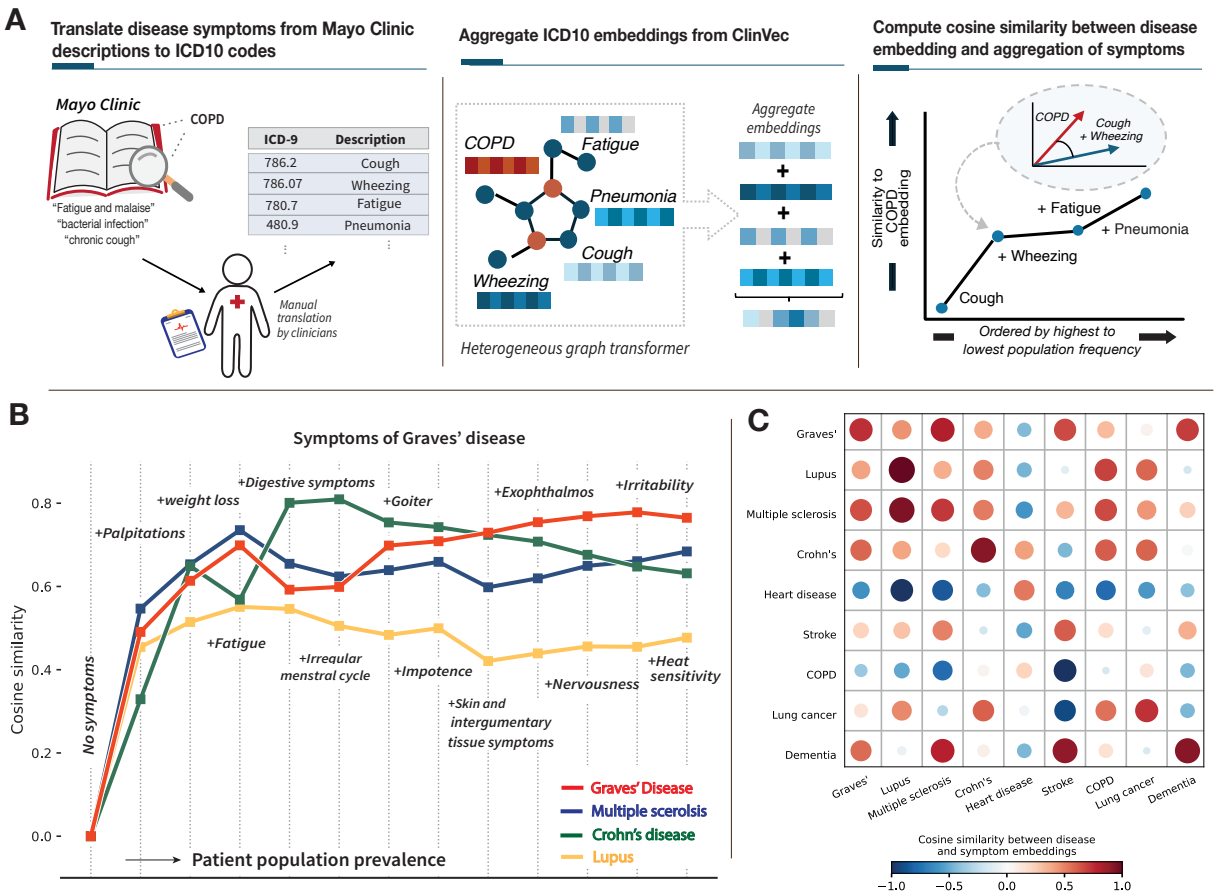

Figure 4: Embedding operations mirror symptomatic manifestations of clinical conditions. **(A)** Overview of the approach for constructing disease symptom embeddings. **(B)** Line plot showing the cosine similarity across four autoimmune diseases and the aggregation of Graves' disease symptoms. Aggregated embeddings are computed by

performing a vector sum across the embeddings for each symptom. **(C)** Heatmap showing cosine similarity between each disease embedding and the aggregation of symptoms. Sizes of circles correspond to the magnitude of the cosine similarity.

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[92] and four are among the most common autoimmune disorders in adult populations[93]. Using clinical descriptions of diseases from the Mayo Clinic[94–98] 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 S5). 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 S7). 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).

After identifying this pattern, we evaluated whether individual symptoms exhibit consistent syntactic and semantic structure within the embedding space. To approximate the disease specificity of each symptom, we ranked symptoms by their frequency in a sample of patients from Clalit Health Services (CHS), from most to least common. For each of the nine diseases, we incrementally aggregated symptom embeddings in order of decreasing frequency. As less common symptoms were added, the combined symptom embedding progressively aligned more closely with the target disease embedding (Supplementary Figures S2, S3). To test the robustness of this trend, we perturbed key components of the embedding aggregation process. Across variations, including swapping the target disease, substituting related symptoms, and introducing randomly selected symptoms to mimic real-world variability, the symptom embedding consistently moved closer to the corresponding disease embedding in the latent space (Supplementary Figure S4).

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

335 compute the initial cosine similarity between the disease embedding for each of the four diseases
336 and the first symptom embedding, "palpitations", (the Graves' symptom found to be most preva-
337 lent in the sampled patient population), but no disease has a cosine similarity score > 0.40 (Fig-
338 ure 3B). Given the relatively high prevalence of this symptom in the general patient population
339 and lack of disease specificity (*freq*=0.232), it is unsurprising that initial symptoms do not show
340 a high correlation with any of the immune conditions. However, as we add symptoms that are
341 relatively less common within the general population, such as 'goiter' (*freq*=0.029), the composi-
342 tion of symptoms becomes increasingly similar to the representation of Graves' disease, and the
343 similarity with the other three autoimmune conditions begins to diverge. These patterns mirror
344 how individual symptoms may not provide sufficient information to assess the probability of a
345 specific condition, but as a patient accumulates more symptoms over time, the aggregation of
346 these symptoms can stratify patients by disease risk.

347 **Clinical knowledge embeddings stratify patients by disease risk and severity**

348 We next investigate how *ClinVec* can be utilized in the context of patient outcome prediction, in-
349 cluding disease progression and disease risk. Using only information derived from *ClinVec*, we
350 compute a disease-specific risk score for various conditions and the relative disease risk for each
351 patient. Analogous to polygenic risk score frameworks[99], we compute a phenotype risk score[100]
352 by scanning a patient's EHR and aggregating the embeddings across conditions with the highest
353 similarity to the disease of interest (Methods). We validated our approach using retrospective
354 EHR data from 4.57 million individuals[101] at CHS and specifically focused on patients listed in
355 the CHS chronic registry, a registry maintained by CHS that monitors individuals with specific
356 chronic conditions[102]. We examine three clinical conditions with the largest sample sizes within
357 the chronic registry: chronic kidney disease (CKD), chronic obstructive pulmonary disease
358 (COPD), and heart failure.

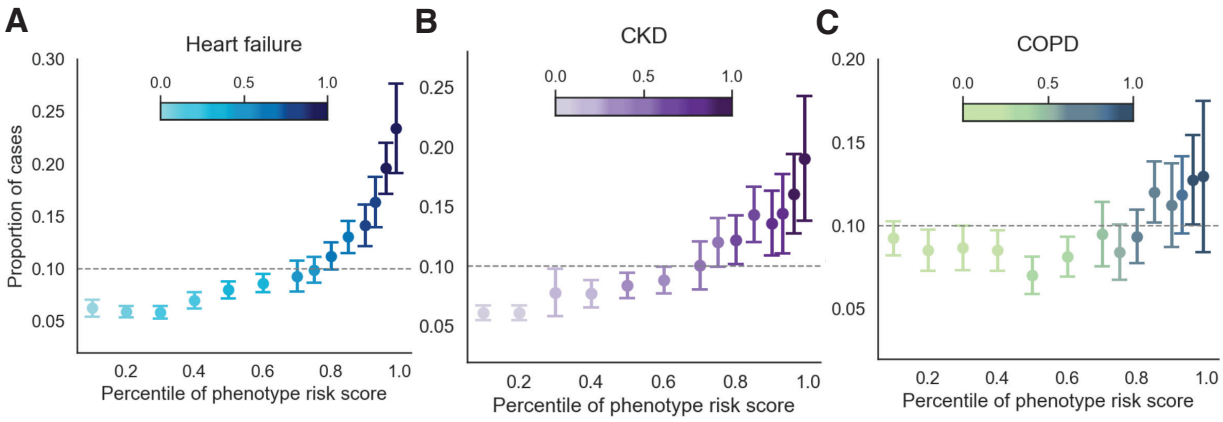

Figure 5: Clinical code embedding risk scores correlate with disease prevalence and severity. (A-C) We provide scatter plots showing the risk score percentiles of each disease versus the disease prevalence within each bin. Dots are shaded according to disease prevalence. Columns are divided by diseases: heart failure, CKD, and COPD.

For each condition, we constructed a matched case-control cohort (Supplementary Table S6, Supplementary Figure S5) and used 5 years of clinical history to compute a relative risk score for each patient. For a target disease (here represented by an ICD10CM embedding), we identify embeddings (features) most relevant to the disease by selecting all codes within a given radius in the embedding space (see Methods). Each patient's score is computed as a weighted sum of these features, where the weight is calculated as the cosine similarity between the feature embedding and target disease embedding. 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[103]. 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.

Quantifying patients' disease risk according to their risk score percentile of the total score distribution, we observe the separation of patients according to disease status across all three conditions, particularly at the extreme end of the risk distribution (Figure 5). Within the group of individuals in the 10th percentile of the risk score distribution for CKD, there is a disease prevalence of 6.2%. For patients above the 99th percentile, the disease prevalence increases to 23.4%. 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. 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.

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 (Supplementary Figure S6). For CKD (age=72 sex=Male), individuals within the top quintile of this score distribution for CKD had a relative 42.6% lower 6-year survival compared to those in the bottom quartile (top-quartile: 56.8.3%, *bottom-quartile*: 76.5%). This trend is particularly remarkable considering the risk scores were inferred without any institution-specific, disease-specific, or patient-level information during training.

ClinVec **provides structured clinical knowledge to improve LLM performance**

Clinical embeddings provide a structured approach to incorporate domain-specific knowledge into foundation models such as LLMs. One common strategy is knowledge injection, where embeddings from external knowledge sources are appended to the LLM input. Recent studies show that this approach improves performance on tasks such as medical question answering by grounding LLM outputs in a clinically relevant context[104–107]. We adapt MMed-Llama[106], an open-source LLM composed of 8 billion parameters. To test whether *ClinVec* embeddings improve LLM performance beyond what can be achieved with domain-specific training data alone, we evaluated our knowledge token injection approach using an LLM already trained on medical content. We used parameter-efficient fine-tuning to integrate 42,883 clinical knowledge tokens into the LLM, where each token corresponded to a *ClinVec* embedding of a UMLS concept (Supplementary Figure S7). This approach allows domain adaptation without retraining the entire model or requiring a labeled clinical dataset. It provides a resource-efficient strategy for incorporating structured clinical knowledge into LLMs.

To assess the impact of clinical knowledge embeddings on LLM performance, we used the MedQA[1–4] dataset, which contains questions from the US Medical Licensing Exams (USMLE).

We applied the UMLS MetaMap API to map question text to UMLS concepts, then retrieved the corresponding *ClinVec* embeddings and appended them to the tokenized prompt (see Methods). This approach incorporates structured, clinically grounded knowledge without retraining the underlying LLM. Instead, *ClinVec* supports retrieval-augmented generation, where embeddings relevant to the input are used to guide the LLM response.

Assessing performance with a held-out test set of MedQA (N=1,247 questions), we find a 10.3% increase in question-answering accuracy when incorporating the clinical knowledge embeddings (acc=0.64) compared to question answering without it (acc=0.58). We observe the largest performance gains when questions are short, likely because the language model has limited contextual information to generate accurate answers. Adding just one or two clinical knowledge embeddings leads to an average improvement of 11.7%, suggesting that the relationship information encoded in these embeddings helps compensate for missing context. Performance continues to improve as more tokens are added, reaching a peak at around 9 to 10 tokens. Beyond this point, additional embeddings introduce less relevant or less specific information, which reduces overall accuracy.

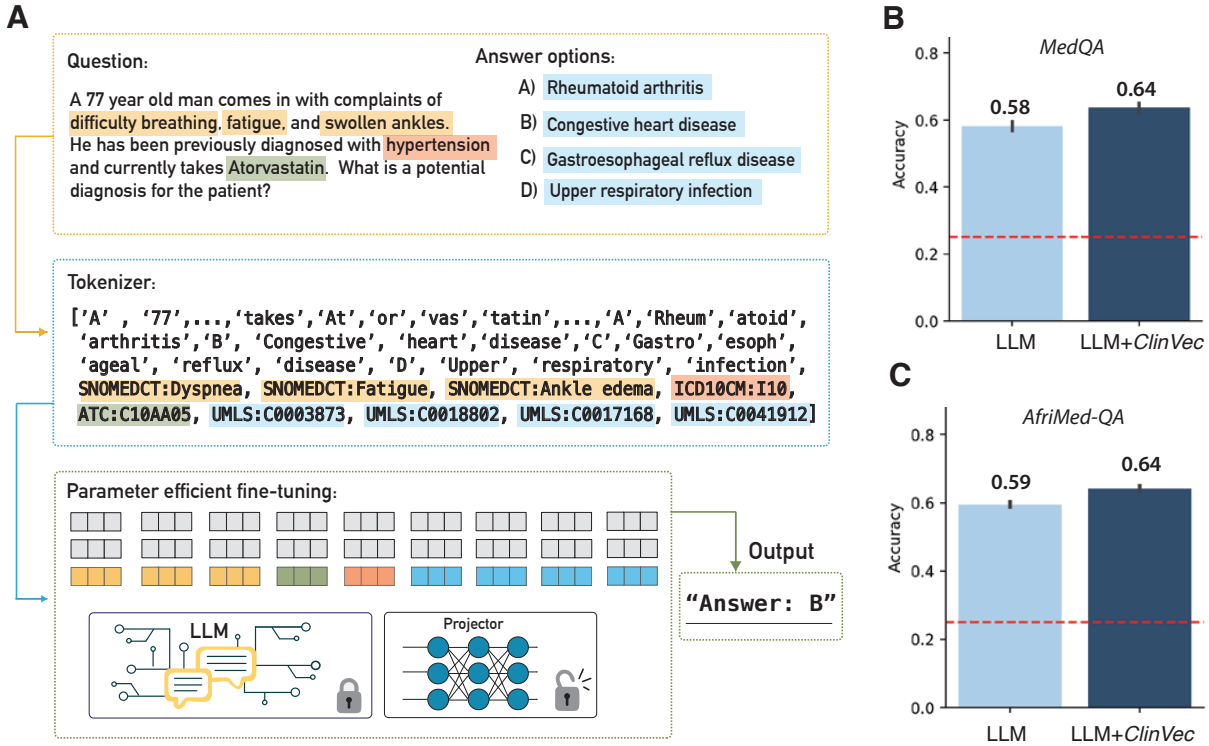

Figure 5: Integrating *ClinVec* with large language models improves medical question-answering. (A) Overview diagram of question-answering with knowledge injection using *ClinVec*. **(B)** We report the accuracy of

N=1,247 medical questions in the MedQA dataset. The red line denotes random performance with four candidate answer options. **(C)** Accuracy of N=3,724 multiple-choice questions from AfriMed-QA dataset. The red line denotes baseline performance with five answer options.

To assess generalizability across different clinical scenarios, we also examined the recently released AfriMed-QA dataset[109], which consists of 3,724 multiple-choice questions collected from healthcare settings across 16 African countries. Since the dataset was released in early 2025, there is no data leakage between the training and fine-tuning datasets. We observe a similar magnitude of performance improvement where we observe accuracy of 0.64 with the addition of clinical knowledge embeddings compared to 0.59 without—a 8.5% increase. Breaking down performance by clinical specialty (a label provided in the original dataset), we identify the specialties with the greatest performance gains, including hematology (+37.0%) and ophthalmology (+20.3%). Given the specificity of these fields, the addition of clinical knowledge yields the most substantial improvements in question-answering performance. Hematology shows the largest gains, likely due to the distinct disease presentations encountered in African settings that are underrepresented in the training data of most large language models. For instance, one correctly answered question concerns antiretroviral treatment regimens for HIV-associated anemia, a clinical scenario more common in regions with high HIV prevalence. This example illustrates how anchoring model responses in structured, domain-specific knowledge from *ClinVec* helps address regional variation and improves accuracy in specialized medical contexts. We observe consistent performance improvements across both U.S.- and Africa-based clinical questions, suggesting that *ClinVec* supports generalizable medical reasoning.

Discussion

We introduce *ClinGraph*, a clinical knowledge graph, and *ClinVec*, a corresponding set of 153,166 clinical vocabulary embeddings, that together translate structured clinical information into a machine-readable format for use in AI models. These embeddings capture patterns among clinical codes that reflect human anatomy, disease presentation, and healthcare organization. They encode semantic relationships that support the combination and comparison of clinical concepts. We demonstrate that the embeddings enable patient-level prediction across diverse conditions without using patient data during training. Evaluations by inter-institutional panels of clinicians from the United States and Israel confirm that the embeddings align with current clinical

knowledge and practice. This resource offers a generalizable framework for sharing clinical knowledge and developing transferable, knowledge-grounded AI models.

Clinical code embeddings offer a way to represent structured clinical knowledge for use in AI models. By encoding relationships among diagnoses, procedures, medications, and laboratory tests, the embeddings create a knowledge-grounded representation of clinical data that can be used in downstream machine learning applications. These representations are well-suited to serve as inputs to foundation models, large-scale models pre-trained on unlabeled data and adapted to specific tasks through fine-tuning with labeled datasets[110]. The foundation model approach has shown success in natural language processing and computer vision, supported in part by the availability of publicly shared models such as DeepSeek[111] and LLaMA[112]. Our resource can be used in retrieval-augmented generation[113] systems where embeddings serve as a structured knowledge base. In this setting, relevant embeddings can be retrieved based on a user prompt to provide clinical context and support more accurate and informed model outputs[41].

While this resource offers a foundation for advancing precision medicine, it has several limitations. First, the current knowledge representations do not model uncertainty. Clinical knowledge is derived from heterogeneous sources, and the frequency or relevance of specific observations can vary across populations. For example, while disease guidelines may list several symptoms, their prevalence and clinical importance often differ between individuals. Second, the quality of the embeddings depends on the accuracy of the underlying KG. Errors such as incorrect or incomplete edges can introduce noise and distort the learned representations. As clinical knowledge evolves, the graph must be regularly updated to reflect current clinical standards and practice. Third, the embeddings do not incorporate contextual information, such as age or sex, which can influence the presentation and interpretation of clinical concepts. For example, the clinical presentation of a heart attack can vary significantly between males and females[114], but this delineation is not considered in the construction of the embeddings.

ClinVec embeddings cover common and chronic diseases, with limited representation of genetic disorders and Mendelian conditions. Future work will incorporate codes from phecodeX[115], an updated vocabulary that includes chromosomal anomalies, as well as resources focused on Mendelian diseases, such as the Online Mendelian Inheritance in Man (OMIM)[116] and Human Phenotype Ontology (HPO)[74]. We also plan to integrate databases that capture biological mechanisms,

including those involving genes and proteins, to improve the representation of disease with complex or poorly defined clinical features.

ClinGraph consolidates fragmented coding systems into a unified representation of clinical knowledge, addressing interoperability challenges while protecting patient privacy. The embeddings are designed to capture clinical knowledge that generalizes across care settings and aligns with established medical practices. Unlike datasets that only reflect recorded health states, *ClinVec* embeddings incorporate information about clinical guidelines, mechanistic associations, and disease-treatment pathways. Because the embeddings are derived from structured vocabularies and not patient-level data, they can be shared across institutions without risk of exposing identifiable information. Unlike models pre-trained on patient records, which may inadvertently retain or memorize sensitive information[117,118], *ClinVec* embeddings can be shared across healthcare systems without risking patient privacy. This approach also reduces dependence on specific patient populations and supports diverse clinical applications. By embedding structured clinical knowledge into AI models, this resource helps ensure that predictions are grounded in established medical understanding[119]. Resources such as unified clinical vocabulary embeddings can strengthen the reliability and generalizability of AI models.

509 **Methods**

510 *Standardized medical vocabularies*

511 We use standardized medical vocabularies to construct all clinical concepts in *ClinGraph* and
512 *ClinVec*, supporting broad applicability across clinical use cases: Unified Medical Language
513 System (UMLS), International Classification of Diseases (ICD)[120], Anatomical Therapeutic
514 Chemical (ATC) Classification[121], RxNorm[122], Systemized Nomenclature of Medicine – Clinical
515 Terms (SNOMED CT)[123], Current Procedural Terminology (CPT)[124], Logical Observation Iden-
516 tifiers Names and Codes (LOINC)[125], and phecodes[126]. Here, we define a clinical code as an en-
517 try within a medical vocabulary with a unique identifier.

518 We begin by integrating the clinical codes and relationships defined in the UMLS vocabulary.
519 We incorporate all UMLS CUIs (concept unique identifier) with a source vocabulary in
520 ICD10CM, SNOMED CT, LOINC, RxNorm, ATC, or CPT (N=42,883 nodes). Next, we inte-
521 grate all ICD10CM codes up to 2 decimal places (N=34,842) and all ICD9CM codes. We also
522 add information from the phecode coding system which represents clinical phenotypes as phe-
523 codes, formed by aggregating groups of diagnostic codes (i.e., ICD codes) into clinical sub-
524 groupings (N=1,817) [127–129]. Phecodes were developed to facilitate the secondary use of EHRs
525 for biomedical research rather than for patient monitoring or hospital administration. Numbers
526 after the decimal point reflect a hierarchical structure similar to the ICD code hierarchy. Under
527 each leaf phecode is the set of corresponding ICD-9/10 codes where each phecode has between 1
528 – 20+ associated ICD-9/10 codes. Next, we add nodes representing LOINC Part codes, which are
529 standardized attribute values used to construct the LOINC codes that specify a given laboratory
530 test. We restrict our inclusion to only LOINC Part component codes which refer to the substance
531 being tested (N=14,399). We also incorporate all CPT codes (N=4,377), ATC codes (levels 3-5:
532 N=5,179), and RxNorm drug groups (N=10,130).

533 Next, we integrate a subset of SNOMED CT codes as nodes in *ClinGraph*. Because the
534 SNOMED CT vocabulary includes many codes (>350,000), we wanted to restrict the inclusion
535 to the most clinically relevant codes. We begin by selecting all SNOMED CT codes from the
536 CORE (Clinical Observations Recording and Encoding) Subset, a set of most frequently used
537 codes provided by UMLS constructed to maximize data interoperability across institutions. Next,
538 we include SNOMED CT codes from the Convergent Medical Terminology subset diagnosis list,

539 a set of core terms originally curated by Kaiser Permanente and now used across numerous
540 healthcare systems.

541 *Relationships between clinical concepts*

542 To construct edges between pairs of clinical concepts (nodes), we first incorporate relationships
543 defined by the PheMap database, an existing knowledge base of biomedical knowledge that de-
544 fines a set of relationships between the codes and various clinical vocabularies[83]. This database
545 offers relationship pairs linking phecodes and clinical codes from other medical vocabularies, in-
546 cluding RxNorm, CPT, LOINC, and SNOMED CT. Relationships defined in PheMap were de-
547 termined through text mining and natural language processing of open-source biomedical litera-
548 ture. These relationships can be interpreted as a phenotype being "associated with" a given clini-
549 cal concept. To construct relationships between different medication codes, we use ATCPROD,
550 a product-level mapping of RxNorm to ATC-4 classes. Edges between ICD9CM and ICD10CM
551 codes are derived from the General equivalence mappings (GEM), the official crosswalk map-
552 ping diagnoses codes to and from ICD9CM and ICD10CM.

553 *Constructing ClinGraph*

554 To efficiently represent clinical knowledge in a format optimized for machine learning, we for-
555 mally organize the clinical codes and relationships, as described above, into a heterogenous
556 knowledge graph[130] (KG) called *ClinGraph.* We draw an undirected edge between two clinical
557 codes if they have a pairwise relationship as described above in the clinical vocabulary aggrega-
558 tion process. Multiple edges are drawn if two nodes have a relationship defined multiple times
559 (e.g., a relationship between a phecode and ICD-9CM code derived from the phecode hierarchy
560 and the PheMap knowledge base). The final graph comprises 153,166 nodes across seven node
561 types and 2,873,025 edges, and the distribution of nodes and edges can be found in Supplemen-
562 tary Table S1, S2.

563 We model 4 distinct node types: diagnosis codes (ICD9CM, ICD10CM, phecodes), medications
564 (ATC, RxNorm), laboratory tests/procedures (LNC, CPT), and clinical concepts/vocabularies
565 (SNOMED CT, UMLS). By grouping similar vocabularies (e.g., RxNorm and ATC), we allow
566 for information sharing across similar nodes. However, modeling distinct edge types would re-
567 quire the harmonization of relationship types across databases. For databases without defined
568 edge type semantics, such as PheMap, this harmonization process would likely require additional

resources and clinical expertise. For this reason, we only model distinct node types and leave the implementation of edge-type specific embeddings as future work.

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 $f_v: \mathcal{V} \rightarrow \mathcal{T}^v$ 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:

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

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	$\mathbb{R}^1$	-
$\boldsymbol{\mathcal{H}_{\mathcal{G}}}$	Graph neural network function	$\mathbb{R}^d \rightarrow \mathbb{R}^d$	Y
$\boldsymbol{x_u}$	Feature vector of node u	$\mathbb{R}^d$	-
$\boldsymbol{z_u}$	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
$\boldsymbol{h_u^{(l)}}$	Embedding of node u at the lth layer	$\mathbb{R}^d$	Y
$\mathbf{ATT}(\boldsymbol{s}, \boldsymbol{t})$	Attention vector for node u	$\mathbb{R}^d$	Y
h	Number of heads	$\mathbb{R}^1$	-
$\boldsymbol{W_Q^i}, \boldsymbol{W_K^i}$	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
$\boldsymbol{W_Q^i}, \boldsymbol{W_K^i}$	Weight matrices for attention computation	$\mathbb{R}^{d \times \frac{d}{h}}$	Y
$\boldsymbol{W_V}$	Weigh matrix for attention computation	$\mathbb{R}^{\frac{d}{h} \times \frac{d}{h}}$	Y

$\boldsymbol{K_s^i}, \boldsymbol{Q_t^i}$	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	-	-
$\Big\Vert_{\boldsymbol{n}}(\boldsymbol{x^i})$	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)[86]. In this work, we briefly describe the key model steps. The complete derivation of HGT is detailed in previous work[86].

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, $\boldsymbol{h}_u^{(l+1)} = \mathrm{UPD}\left(\boldsymbol{m}_t^{(l)}, \boldsymbol{h}_u^{(l)}\right)$ where $\boldsymbol{h}_t^{(l+1)}$is the embedding of node t at the *l+1* layer and UPD(·) 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:

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

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 0th 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.

Multi-head attention mechanism

The HGT uses multi-head attention, which is a key feature of transformer architecture[87]. 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:

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

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

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

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

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

The input vector is $\boldsymbol{h}_s^{(l-1)} \in \mathbb{R}^{\boldsymbol{d}\times 1}$, 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:

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

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

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

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,

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

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[88,131–133] and is formulated as follows:

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

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

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 *ClinGraph*, 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 *ClinGraph* 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 *ClinGraph* 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 2,873,025 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 *ClinGraph*. 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).

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

723 validate this, we perform a human pilot evaluation with groups of clinical experts from Brigham
724 Women's (N=1), Mass General (N=1), Boston Children's Hospital (N=1), and CHS (N=3).

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

737 Clinicians were provided with shuffled disease lists of case and control codes. For a given dis-
738 ease, a single clinician evaluated the full list of codes. They were provided a brief background of
739 SNOMED CT codes and a scoring rubric where they were asked to evaluate each SNOMED CT
740 code according to how related each code is to the given disease phenotype. We utilize a Likert
741 scale[134] based on scores 1 – 5: 'unrelated' (0) and ' highly related' (5). If a code is too broad or
742 non-specific (e.g. 'Clinical evaluation'), these are assigned a 0.

743 **Visualizing the latent space**

744 To visualize the *ClinVec* embedding space, we perform UMAP[135] directly on the embedding vec-
745 tors and plot the first two dimensions. When performing dimensionality reduction, we use phe-
746 codes, ICD10CM, ATC, LOINC, CPT, and SNOMED CT embeddings. Phecode categorizations
747 are pre-defined from the phecode v1.2 mapping[136]. To visualize the latent space of phecodes
748 within the digestive system, we performed UMAP only on the subset of ICD10CM embeddings
749 within the 'Digestive' category. We highlight ICD10CM codes broadly corresponding to
750 mouth/teeth conditions (phecodes 520-530), functional digestive disorders (560-570), and biliary
751 tract disorders (570-580). To highlight embeddings related to specific diseases, we compute the
752 cosine similarity between an ICD10CM embedding (representing a disease) and each embedding
753 within the latent space. We visualize this relationship by shading each embedding within the

754 UMAP space according to the cosine similarity. We highlight relevant codes for lupus and type 1
755 diabetes to provide additional interpretation of the latent space map. This information was de-
756 rived from the guidelines provided by the National Resource Center on Lupus[137] and the Centers
757 for Disease Control and Prevention[138].

758 *Comparing embeddings from PrimeKG and SNOMED-CT KG*

759 We generate embeddings for PrimeKG and SNOMED-CT KG using the same graph transformer
760 architecture and hyperparameters as *ClinGraph*. For PrimeKG, we modeled 10 node types as de-
761 fined by the KG. For SNOMED-CT KG, we only used a single node type (i.e. homogenous
762 graph) since all nodes represent SNOMED CT codes. We trained both models for 5 epochs and
763 generated embeddings of size 128 for all SNOMED CT codes in SNOMED-CT KG and all dis-
764 eases (MONDO, HPO nodes) and drugs (DrugBank nodes) for PrimeKG. Clustering metrics, in-
765 cluding silhouette score and adjusted mutual information, were computed according to ATC
766 level 1 and phecode category.

767 Because neither KG have ICD or ATC codes, we mapped the nodes as follows. For SNOMED-
768 CT KG, we mapped each SNOMED CT code to its UMLS CUI and then to its corresponding
769 ICD10CM code if present. If a UMLS CUI does not have a corresponding source ICD10CM
770 code, we removed the node embedding from the cluster calculation. To map SNOMED CT to
771 ATC, we first mapped SNOMED CT codes to its UMLS CUI. We then used UMLS to identify
772 UMLS CUI codes with 'ingredient-of' relationships that describe the active ingredient (repre-
773 sented as ATC) present in a given UMLS CUI drug. SNOMED CT codes that did not have a
774 mapped ATC ingredient were removed from the clustering calculation. Clustering metrics were
775 computed using all 128 dimensions from each set of embeddings, and cluster visualizations were
776 produced using UMAP.

777 *Zero-shot embedding retrieval tasks*

778 We downloaded embeddings from snomed2vec
779 (https://github.com/KhushbuAgarwal/Snomed2Vec) and cui2vec (https://github.com/beaman-
780 drew/cui2vec) directly from the original publications. We formulated a set of 12 retrieval tasks
781 using well-established disease-symptom and disease-indication lists (Supplementary Table S8).
782 Top ranked neighbors are determined using cosine similarity between embeddings. We report

783 each model's performance using 'hits @ k' which measures the proportion of diseases where a
784 relevant symptom/drug is ranked among the top k items within the embedding space.

785 **Disease embedding arithmetic**

786 *Constructing disease symptom embeddings*

787 To demonstrate the syntactic and semantic properties of *ClinVec*, we use vector-level arithmetic
788 to aggregate together disease symptoms to form a 'disease symptom embedding'. Aggregation is
789 performed using sum pooling, but any vector-level aggregation method could be used in practice.
790 For a given disease, we retrieve the list of symptoms from the Mayo Clinic Disease Description
791 Guidelines[139]. A clinician manually translates each symptom into its corresponding ICD10CM
792 code. We provide these translations for all nine diseases discussed within this analysis (Supple-
793 mentary Table S5). To demonstrate the trend of increased similarity between a disease embed-
794 ding and the disease symptom embedding, we order symptoms from highest to lowest preva-
795 lence. We approximate population prevalence using a random sample of 200,000 patients from
796 the CHS database.

797 Disease symptom embeddings are first initialized with a vector of zeros. We then add the embed-
798 dings from all symptoms (represented by an ICD10CM embedding). For visualization, we com-
799 pute the cosine similarity between the disease symptom embedding at each step and the target
800 disease embedding, represented using an ICD10CM embedding. Across diseases, we see a broad
801 trend of increased similarity with the disease embedding as more symptoms are aggregated.

802 *Robustness analysis*

803 To assess the robustness and generalizability of these trends, we perform the same analysis under
804 various conditions while perturbing different variables in the framework. This includes analyses
805 assessing the uncertainty in selecting disease symptoms, incorrect disease specification, and the
806 impact of random symptoms. First, due to the noise and uncertainty associated with using diag-
807 nosis codes as proxies for disease symptoms[13], we tried to model real-world circumstances by
808 varying the ICD10CM code used to represent each symptom. We selected a set of candidate
809 ICD10CM codes for each symptom by taking the original code from the clinician review and se-
810 lecting all ICD10CM codes with the same parent code by removing one decimal place. For ex-
811 ample, for code '786.2' (Cough), the candidate set would include '786.1' (Stridor), '786.0'

812 (Dyspnea and respiratory abnormalities), etc. Then, when selecting each symptom for aggrega-
813 tion, a random code from the candidate set is selected. This process is repeated 10 times, yielding
814 similarity curves constructed from 10 different disease symptoms.

815 Next, we wanted to simulate when the disease of interest is mis-specified compared to the symp-
816 tom list. This reflects situations where the symptoms do not describe the target disease of interest
817 but, instead, a closely related condition. To imitate this, we selected a set of diseases similar to
818 the target disease; this was done by looking at other ICD10CM with the same parent code. When
819 computing the similarity between symptom embedding and disease embedding composition, we
820 used the phecode embedding of the alternative disease. Finally, we wanted to assess the situation
821 where incorrect information is included in the symptom list. We simulated this by adding a ran-
822 domly selected ICD10CM code to the symptom list and included it in the computation of the ag-
823 gregated symptoms. Because we use vector addition to aggregate symptoms and addition is com-
824 mutative, we place the random symptom at the beginning of the aggregation procedure.

825 **Phenotype risk score**

826 *Case/control construction*

827 To define cases, we first selected all patients with at least one ICD9CM code from the list of in-
828 clusion codes (CHS encodes their EHR using ICD9CM). We then filter patients to include only
829 patients listed on the CHS chronic registry to confirm they have received the disease diagnosis.
830 Because patients are not guaranteed to be added to the chronic registry when they are first diag-
831 nosed, we treat the first occurrence of the inclusion ICD9CM code as the date of diagnosis. For
832 each patient, we set the index date as January 1 of the year of diagnosis, where only information
833 occurring before the index date will be included in the risk score computation. For prediction, we
834 select all EHRs recorded within the past 5 years of the index date.

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

842 least 1 diagnosis in the EHR). The complete case-control process is illustrated in Supplementary
843 Figure S5.

844 *Risk score computation*

845 For a given disease, a risk score is computed as the weighted sum of relevant clinical codes,
846 where the weight is the cosine similarity between the embedding of each code and the disease
847 embedding. We restricted the set of potential features to laboratory tests, medications, and
848 ICD9CM codes with at least a 0.1% prevalence within the CHS patient population. The CHS
849 EHR did not have SNOMED CT or CPT codes for analysis. To determine the clinical codes re-
850 lated to each disease, we compute the cosine similarity between all clinical codes within the la-
851 tent space and the target disease embedding, represented by a phecode. The final set of features
852 is the top k codes, where k=20, and the feature weights are the cosine similarity scores. In prac-
853 tice, the choice of k can be treated as a hyper-parameter and optimized with a validation dataset,
854 similar to the p-value thresholding of polygenic risk scores. Patient risk scores are computed as a
855 weighted sum where the associated weight is added to the sum if the patient has the given code
856 in their EHR and given a 0 otherwise. Features are counted once in the score computation, even
857 if a patient receives the codes multiple times.

858 *Survival analysis*

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

866 **Integrating *ClinVec* with LLMs**

867 We augment MMed-Llama[106]—an 8-billion-parameter open-source medical foundation model
868 that already incorporates domain-specific knowledge through pre-training—with 42,883 *ClinVec*
869 embeddings and a 4-layer MLP projector module to align embedding dimensions which is
870 trained with parameter efficient fine-tuning. We download the pre-trained model from Hugging

Face (https://huggingface.co/Henrychur/MMed-Llama-3-8B) and use the authors' fine-tuning pipeline to perform parameter efficient fine-tuning with LoRA[140] (https://github.com/MAGIC-AI4Med/MMedLM/tree/main). We used the same prompts and sets of hyper-parameters provided by MMed-Llama, specifically tuning the following target modules: 'q_proj', 'v_proj', 'v_proj', 'mlp'. Fine-tuning was performed on the training English subset of multiple-choice questions from the MedQA dataset (https://huggingface.co/datasets/VodLM/medqa).

MetaMap

We annotate each multiple-choice question using the MetaMap[141] tool, which annotates natural language text with UMLS CUI codes. For each question, we concatenated the text from the question and the answer options and submitted the text to the MetaMapLite REST API (https://ii.nlm.nih.gov/metamaplite/rest). We limit the choice of UMLS CUI codes to the following semantic types: "Sign or Symptom", "Disease or Syndrome", "Pharmacological Substance", "Mental or Behavioral Dysfunction", "Chemical", "Anatomical Structure".

AfriMed-QA

We also perform a validation study on the AfriMed-QA[109] dataset (https://huggingface.co/datasets/intronhealth/afrimedqa_v2). We restricted to only multiple-choice questions and removed questions that had more than one assigned correct answer. The original dataset provides clinical subspecialty information for each question.

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 *ClinGraph* knowledge graph, are available via Harvard Dataverse https://doi.org/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, Ayush Noori, Intae Moon, and Shvat Messica 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).

955 12. O'Malley, K. J. *et al.* Measuring Diagnoses: ICD Code Accuracy. *Health Serv. Res.* **40**,
956 1620 (2005).
957 13. Botsis, T., Hartvigsen, G., Chen, F. & Weng, C. Secondary Use of EHR: Data Quality Is-
958 sues and Informatics Opportunities. *Summit Transl. Bioinforma.* **2010**, 1–5 (2010).
959 14. Sudat, S. E., Robinson, S. C., Mudiganti, S., Mani, A. & Pressman, A. R. Mind the clini-
960 cal-analytic gap: Electronic health records and COVID-19 pandemic response. *J. Biomed. In-*
961 *form.* **116**, 103715 (2021).
962 15. Binkheder, S. *et al.* Real-World Evidence of COVID-19 Patients' Data Quality in the
963 Electronic Health Records. *Healthcare* **9**, 1648 (2021).
964 16. Chalmers, R. J. G. Health Care Terminology for the Electronic Era. *Mayo Clin. Proc.* **81**,
965 729–731 (2006).
966 17. Song, Z. *et al.* Physician Practice Pattern Variations in Common Clinical Scenarios
967 Within 5 US Metropolitan Areas. *JAMA Health Forum* **3**, e214698 (2022).
968 18. Baicker, K., Chandra, A., Skinner, J. S. & Wennberg, J. E. Who you are and where you
969 live: how race and geography affect the treatment of medicare beneficiaries. *Health Aff. Proj.*
970 *Hope* **Suppl Variation**, VAR33-44 (2004).
971 19. Baicker, K., Chandra, A. & Skinner, J. S. Geographic variation in health care and the
972 problem of measuring racial disparities. *Perspect. Biol. Med.* **48**, S42-53 (2005).
973 20. Medicine, I. of, Services, B. on H. C. & Programs, C. on R. H. I. P. M., Payment, and
974 Performance Improvement. *Rewarding Provider Performance: Aligning Incentives in Medi-*
975 *care*. (National Academies Press, 2007).
976 21. Blumenthal, D. & Tavenner, M. The "Meaningful Use" Regulation for Electronic Health
977 Records. *N. Engl. J. Med.* **363**, 501–504 (2010).
978 22. McIlvennan, C. K., Eapen, Z. J. & Allen, L. A. Hospital Readmissions Reduction Pro-
979 gram. *Circulation* **131**, 1796–1803 (2015).

980 23. Morden, N. E., Colla, C. H., Sequist, T. D. & Rosenthal, M. B. Choosing Wisely — The
981 Politics and Economics of Labeling Low-Value Services. *N. Engl. J. Med.* **370**, 589–592
982 (2014).
983 24. Fisher, E. S. *et al.* The implications of regional variations in Medicare spending. Part 1:
984 the content, quality, and accessibility of care. *Ann. Intern. Med.* **138**, 273–287 (2003).
985 25. Charlesworth, C. J., Meath, T. H. A., Schwartz, A. L. & McConnell, K. J. Comparison of
986 Low-Value Care in Medicaid vs Commercially Insured Populations. *JAMA Intern. Med.* **176**,
987 998–1004 (2016).
988 26. Baicker, K., Chernew, M. E. & Robbins, J. A. The spillover effects of Medicare managed
989 care: Medicare Advantage and hospital utilization. *J. Health Econ.* **32**, 1289–1300 (2013).
990 27. Ancker, J. S. *et al.* How is the electronic health record being used? Use of EHR data to
991 assess physician-level variability in technology use. *J. Am. Med. Inform. Assoc.* **21**, 1001–
992 1008 (2014).
993 28. Roland, M. Linking Physicians' Pay to the Quality of Care — A Major Experiment in the
994 United Kingdom. *N. Engl. J. Med.* **351**, 1448–1454 (2004).
995 29. Sirovich, B. E., Gottlieb, D. J., Welch, H. G. & Fisher, E. S. Variation in the Tendency of
996 Primary Care Physicians to Intervene. *Arch. Intern. Med.* **165**, 2252–2256 (2005).
997 30. Berner, E. S., Kasiraman, R. K., Yu, F., Ray, M. N. & Houston, T. K. Data quality in the
998 outpatient setting: impact on clinical decision support systems. *AMIA Annu. Symp. Proc.*
999 *AMIA Symp.* 41–45 (2005).
1000 31. Verheij, R. A., Curcin, V., Delaney, B. C. & McGilchrist, M. M. Possible Sources of Bias
1001 in Primary Care Electronic Health Record Data Use and Reuse. *J. Med. Internet Res.* **20**, e185
1002 (2018).

1003 32. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to
1004 processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479
1005 (2018).
1006 33. Tang, P. C., Ralston, M., Arrigotti, M. F., Qureshi, L. & Graham, J. Comparison of Meth-
1007 odologies for Calculating Quality Measures Based on Administrative Data versus Clinical
1008 Data from an Electronic Health Record System: Implications for Performance Measures. *J.*
1009 *Am. Med. Inform. Assoc.* **14**, 10–15 (2007).
1010 34. Hogan, W. R. & Wagner, M. M. Accuracy of Data in Computer-based Patient Records. *J.*
1011 *Am. Med. Inform. Assoc.* **4**, 342 (1997).
1012 35. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J.*
1013 *Am. Med. Inform. Assoc. JAMIA* **20**, 117 (2012).
1014 36. Williams, R., Kontopantelis, E., Buchan, I. & Peek, N. Clinical code set engineering for
1015 reusing EHR data for research: A review. *J. Biomed. Inform.* **70**, 1–13 (2017).
1016 37. Delvaux, N. *et al.* Coding Systems for Clinical Decision Support: Theoretical and Real-
1017 World Comparative Analysis. *JMIR Form. Res.* **4**, e16094 (2020).
1018 38. Odigie, E. *et al.* Fast Healthcare Interoperability Resources, Clinical Quality Language,
1019 and Systematized Nomenclature of Medicine—Clinical Terms in Representing Clinical Evi-
1020 dence Logic Statements for the Use of Imaging Procedures: Descriptive Study. *JMIR Med. In-*
1021 *form.* **7**, e13590 (2019).
1022 39. Kokkinakis, D. What is the Coverage of SNOMED CT®on Scientific Medical Cor-
1023 pora? in *User Centred Networked Health Care* 814–818 (IOS Press, 2011). doi:10.3233/978-
1024 1-60750-806-9-814.
1025 40. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care.
1026 *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.

1052 52. ACMG recommendations for standards for interpretation of sequence variations. *Genet.*
1053 *Med.* **2**, 302–303 (2000).
1054 53. Poonacha, T. K. & Go, R. S. Scientific evidence underlying National Comprehensive
1055 Cancer Network Clinical Practice Guidelines. *J. Clin. Oncol.* **28**, 6020–6020 (2010).
1056 54. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical
1057 Practice Guidelines. *Clinical Practice Guidelines We Can Trust*. (National Academies Press
1058 (US), Washington (DC), 2011).
1059 55. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph
1060 convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).
1061 56. Nelson, W. *et al.* To Embed or Not: Network Embedding as a Paradigm in Computational
1062 Biology. *Front. Genet.* **10**, (2019).
1063 57. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from mate-
1064 rials science literature. *Nature* **571**, 95–98 (2019).
1065 58. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Princi-
1066 ples, practice, and opportunities. *Inf. Fusion* **50**, 71–91 (2019).
1067 59. Zhou, H. *et al.* Knowledge-guided convolutional networks for chemical-disease relation
1068 extraction. *BMC Bioinformatics* **20**, 260 (2019).
1069 60. Nunes, S., Sousa, R. T. & Pesquita, C. Multi-domain knowledge graph embeddings for
1070 gene-disease association prediction. *J. Biomed. Semant.* **14**, 11 (2023).
1071 61. Informatics, O. H. D. S. and. *Chapter 4 The Common Data Model | The Book of OHDSI*.
1072 62. Himmelstein, D. S. & Baranzini, S. E. Heterogeneous Network Edge Prediction: A Data
1073 Integration Approach to Prioritize Disease-Associated Genes. *PLOS Comput. Biol.* **11**,
1074 e1004259 (2015).
1075 63. Morris, J. H. *et al.* The scalable precision medicine open knowledge engine (SPOKE): a
1076 massive knowledge graph of biomedical information. *Bioinformatics* **39**, btad080 (2023).

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

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

66. 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).

67. 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).

68. 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).

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

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

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

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

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

74. 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).

75. 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).

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

77. 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.

78. 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).

79. 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).

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

81. 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).

82. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).

83. 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).

84. Königs, C., Friedrichs, M. & Dietrich, T. The heterogeneous pharmacological medical biochemical network PharMeBINet. *Sci. Data* **9**, 393 (2022).

85. Alber, D. A. *et al.* Medical large language models are vulnerable to data-poisoning attacks. *Nat. Med.* **31**, 618–626 (2025).

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

87. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (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. Agarwal, K. *et al.* Snomed2Vec: Random Walk and Poincaré Embeddings of a Clinical Knowledge Base for Healthcare Analytics. Preprint at https://doi.org/10.48550/arXiv.1907.08650 (2019).

90. Beam, A. L. *et al.* Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. Preprint at https://doi.org/10.48550/arXiv.1804.01486 (2019).

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

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

93. 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).

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

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

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

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

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

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

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

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

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

103. 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).

104. Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).

105. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

106. Qiu, P. *et al.* Towards Building Multilingual Language Model for Medicine. Preprint at https://doi.org/10.48550/arXiv.2402.13963 (2024).

107. Maharjan, J. *et al.* OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci. Rep.* **14**, 14156 (2024).

108. Jin, D. *et al.* What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. Preprint at https://doi.org/10.48550/arXiv.2009.13081 (2020).

109. Olatunji, T. *et al.* AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset. Preprint at https://doi.org/10.48550/arXiv.2411.15640 (2025).

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

111. DeepSeek-AI *et al.* DeepSeek-V3 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2412.19437 (2025).

112. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).

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

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

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

116. 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).

117. 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.

118. 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.

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

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

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

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

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

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

125. 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).

126. 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).

127. 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).

128. 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).

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

130. 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).

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

132. 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).

133. 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).

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

135. 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).

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

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

138. CDC. Type 1 Diabetes. *Diabetes* https://www.cdc.gov/diabetes/about/about-type-1-diabetes.html (2025).

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

140. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).

1245 141. Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the
1246 MetaMap program. *Proc. AMIA Symp.* 17–21 (2001).

1247