A two-method framework for aligning medical terminologies

Stan Ostaszewski[1], Ömer Durukan Kılıç[1,*], Ensar Emir Erol[1], Michel Dumontier[1] and Remzi Celebi[1]

[1]Maastricht University, Department of Advanced Computing Sciences, Institute of Data Science, Paul-Henri Spaaklaan 1, 6221 GS, Maastricht, Limburg, The Netherlands

Abstract

The healthcare data interoperability relies on aligning disparate terminologies to a unified ontology, such as SNOMED CT. In this work, we present a framework consisting of two methods that leverages expert-curated reference sets for direct 1-to-1 mappings; augmented by algorithmic strategies for incomplete or 1-to-M cases. Firstly, for unmapped procedure (CPT) and medication (NDC) codes, we power logistic regression-based imputation and non-contextual ranking with medBERT text embeddings. Second, for aligning ambiguous diagnoses (ICD-9) codes, a patient-context-aware ranking exploits SNOMED CT's hierarchical structure via three proximity metrics: textual cosine similarity, exact shortest-path distances, and Node2Vec embeddings. Text-based imputation yields high AUCs (0.95+ for NDC, 0.85 for CPT); and contextual ranking with Node2Vec for ICD-9's generic mappings achieve Hits@1 of 0.37-0.45 and Hits@5 of 0.82-0.88 on curated EHR-like tests, outperforming text-only methods while delivering a significant cost reduction over computing exact distances. Interpretable confidences (κ_i) are calculated to reward context-specific outliers, enabling robust entity resolution in graph-triple pipelines. With this framework, we bridge gaps between medical terminologies, reducing mapping ambiguity and increasing interoperability. By supporting expert-curated reference sets with algorithmic and statistical methods, our framework advances scalable semantic integration within the healthcare domain.

Keywords

Entity Alignment, Medical Terminologies, Knowledge Graphs, Embeddings, Contextual Ranking, Electronic Health Records

1. Introduction

The standards used for representing health data vary greatly among coding systems. These different medical terminologies often contain complementary or overlapping information about the same medical concepts [1]. However, identical entities are frequently represented using different identifiers, names, or attribute sets due to variations in data collection practices, ontological standards, languages, abbreviations, and notational conventions. Such heterogeneity poses a major challenge for integrating health data coming from different sources [2].

SeWebMeDA-2026: 9th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, May 10, 2026, Dubrovnik, Croatia

*Corresponding author.

s.ostaszewski@student.maastrichtuniversity.nl (S. Ostaszewski); omer.kilic@maastrichtuniversity.nl (Ö. D. Kılıç); ensar.erol@maastrichtuniversity.nl (E. E. Erol); michel.dumontier@maastrichtuniversity.nl (M. Dumontier); remzi.celebi@maastrichtuniversity.nl (R. Celebi)

https://github.com/diamenciarz (S. Ostaszewski); https://github.com/omerdurukankilic (Ö. D. Kılıç); https://github.com/ensaremirerol (E. E. Erol); http://dumontierlab.com (M. Dumontier); https://github.com/rcelebi (R. Celebi)

0009-0009-2605-7834 (S. Ostaszewski); 0009-0000-7823-4753 (Ö. D. Kılıç); 0000-0002-5739-8860 (E. E. Erol); 0000-0001-7769-4272 (R. Celebi)

The AIDAVA project [3] aims to integrate patient health data collected from different systems into a personal knowledge graph structure and to develop solutions for the interoperability issues that arise in this context. One of the main interoperability issues when integrating patient data arises from the semantic ambiguity created when they are represented using heterogeneous medical terminologies, which complicates the process of accurately aligning and interpreting concepts across systems [4]. Entity alignment, also known as ontology matching or entity resolution, aims to identify entities across different knowledge bases that refer to the same real-world object, as well as to distinguish those that are genuinely distinct. Successful alignment enables semantically richer, more complete, and interoperable knowledge graphs. To allow such solutions, many methods for aligning terms and ontologies are proposed [5]. In this study, we present a methodology for converting terminologies; namely, CPT[1], NDC[2], and ICD-9[3] into their SNOMED CT[4] equivalents to achieve shared semantics. We accomplish this through a two-method alignment framework where two approaches for different terminological challenges are developed.

Our first method finds the best SNOMED CT code match for a medical term by analyzing the text and structure of the coding system itself. We use this for procedures (CPT) and medications (NDC) because many of these codes are missing translations and need an unsupervised method to find a match. This is especially helpful as no expert-approved (i.e. gold standard) mappings readily exist for both terminologies.

Our second method chooses the right SNOMED CT code by looking at a patient's medical history. This is crucial for broad and ambiguous diagnoses, like those in ICD-9. The idea to use patient history came from our early tests with LOINC lab codes. We noticed that experts often translate one LOINC into several SNOMED CT codes to capture the whole meaning. However, this often includes generic "filler" codes that do not add medical value and just create noise. For example the code for the "Qualifier value (362981000)" and "Observable entity (363787002)" are included in almost all translations become unnecessary during downstream training tasks. By filtering out these unhelpful codes, we got cleaner translations during our initial attempts. Applying a similar logic to ambiguous ICD-9 diagnoses, which often map to many different SNOMED CT codes, we decided to use the surrounding medical context of the patient to pinpoint the exact, most accurate translation.

2. Related Work

Entity alignment between medical terminologies is a highly relevant and complex challenge in healthcare data interoperability. To systematically assess and compare the growing number of ontology matching technologies, the Ontology Alignment Evaluation Initiative[5] (OAEI) organizes annual evaluation campaigns. They provide standard test cases and benchmarks to help the community openly compare algorithms. Among the many systems evaluated over the years, tools like LogMap [6] and Matcha-DL[6] have consistently demonstrated strong performance across various tracks and complex alignment tasks.

[1]https://www.ama-assn.org/practice-management/cpt/cpt-overview-and-code-approval
[2]https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory
[3]https://www.who.int/standards/classifications/classification-of-diseases
[4]https://www.snomed.org/
[5]https://oaei.ontologymatching.org/
[6]https://github.com/liseda-lab/Matcha-DL

Beyond these comprehensive matching tools, recent advances have focused specifically on embedding models. When aligning entities across different knowledge graphs, existing embedding models generally fall into two primary categories: translation-based and aggregation-based approaches. Translation-based methods typically focus on mapping elements directly from one graph to another. They operate much like a translator finding the equivalent of a word between two languages. On the other hand, aggregation-based models look at the broader environment of an entity. They function by gathering information from connected nodes, similar to understanding a person by looking at their close group of friends, to create a comprehensive profile before attempting an alignment.

2.1. Translation-based entity alignment

Early efforts in this domain focused on learning how to map entities directly between spaces. A model in this area, MTransE [7], operates by studying the structures of individual knowledge graphs and simultaneously figuring out how to translate their embeddings to match each other. As it is a supervised model, it requires a predefined list of already aligned entity pairs to learn effectively. Building on this foundation, JAPE [8] brings the structures of the input graphs into a single, unified mathematical space. Instead of just looking at the network connections, JAPE also uses shared attributes (such as an entity's age or physical location) to improve the alignment process.

To reduce the reliance on a large set of pre-aligned pairs, researchers developed semi-supervised approaches like BootEA [9]. This model bootstraps its learning by making educated guesses about new alignments and iteratively using those guesses to train itself further. While this method achieves higher accuracy than its predecessors, the repeated guessing process makes the model computationally slower. Other models incorporate rich textual data to assist the translation process. For example, KDCoE [10] is another semi-supervised model that reads through entity descriptions using a specialized recurrent neural network to better align cross-lingual data. Advancing the use of text even further, BERT-INT [11] utilizes the BERT language model [12] to represent not just structural connections, but the actual names, descriptions, and specific values attached to each entity, allowing it to capture complex interactions between them.

2.2. Aggregation-based entity alignment

Instead of translating single entities in isolation, aggregation models create representations by pulling in information from an entity's surrounding network. A prominent example is RDGCN [13], which adapts graph convolutional networks to better understand different types of relationships. It achieves this by constructing a secondary dual graph that maps out how relations themselves interact, integrating this broader contextual view with the original graph data. Offering a more generalized approach, RREA [14] unites several previous alignment methods into a single framework. It improves overall accuracy by generating unique, relation-specific representations for each entity.

Subsequent models have focused on refining how this neighborhood information is gathered and weighed. The NMN model [15] acts as a direct structural upgrade to RDGCN by carefully sampling an entity's immediate, one-hop neighbors. It uses an attention mechanism, which is akin to the model deciding which neighbors are the most relevant, to classify individual entities based on the shape of their local network. Finally, recognizing that an entity is defined by more than just its connections, AttrGNN [16] introduces a novel architecture. It successfully integrates standard relationship links with specific attribute data, ensuring that the model evaluates both who an entity is connected to and

what specific traits it possesses.

3. Methodology

Our approach[7] is an alignment framework consisting of multiple methods for different coding challenges. We limit the scope of this study to several disparate medical terminologies, e.g. CPT, NDC, and ICD-9, which will be aligned with the target ontology: SNOMED CT. As a partial ground truth, or prior knowledge, we leverages publicly available reference sets (refsets) curated by domain experts. These are also supported by algorithmic rankings derived from contextual graph embeddings (1-to-M mappings) and text embeddings (incomplete mappings).

3.1. Datasets & Preprocessing

Our alignment methodology relies on diverse datasets drawn from established clinical coding systems. To ensure our computational models run efficiently, we restrict potential SNOMED CT translations for unmapped codes to semantically relevant categories, e.g. we only look for diagnosis entities' translations in the under SNOMED CT's "finding" class' hierarchy. The following outlines the systems we are translating:

1. **SNOMED CT** (Target): "Systematized Nomenclature of Medicine–Clinical Terms" is a massive, highly structured web of medical knowledge. It is also an ontology, i.e. a hierarchical graph where concepts are linked by relationships. An example relationship is: "Viral pneumonia" *is-a* "Infectious disease" and is *caused-by* a "Virus". Comprising over 470,800 codes, it serves as our target structure[8].
2. **CPT** (Procedures): "Current Procedural Terminology" codes are used for billing medical operations. The challenge here is missing translations: out of 11,500 CPT codes, only 3,000 have expert-verified 1-to-1 mappings to SNOMED CT. We must algorithmically find matches for the remaining 8,420 codes among 100,000 potential SNOMED CT procedure concepts. The SNOMED CT candidate subset is selected as the "Procedure (71388002)" concept and its children.
3. **NDC** (Medications): "National Drug Codes" identify specific medications (e.g. 0093-2263-01 for a specific brand of amoxicillin capsules). Because different brands sell the exact same chemical substance, many NDCs translate to the same underlying SNOMED CT concept. Out of 950,000 NDCs, roughly 425,000 remain unmapped and require algorithmic matching against 50,000 SNOMED drug concepts. The SNOMED CT candidate subset is selected as the "Substance (105590001)" concept and its children.
4. **ICD-9** (Diagnoses): "The International Classification of Diseases, Ninth Revision" is used mostly used for billing purposes as well. The main challenge with ICD-9 is ambiguity. While some codes map cleanly 1-to-1, many are broad and translate to a multitude of highly specific SNOMED candidates. For instance, a generic ICD-9 code could correspond to SNOMED CT codes ranging from 2 to 1,436 possibilities (see Figure 1 for a 1-to-many example). We must decide which specific translation is correct for a given patient. To accomplish this, we use curated patient histories from MIMIC-III [17] where diagnoses are also coded using ICD-9 terminology.

[7]Public Repository: https://github.com/AIDAVA-DEV/entity-alignment-public
[8]The license restictions were handled

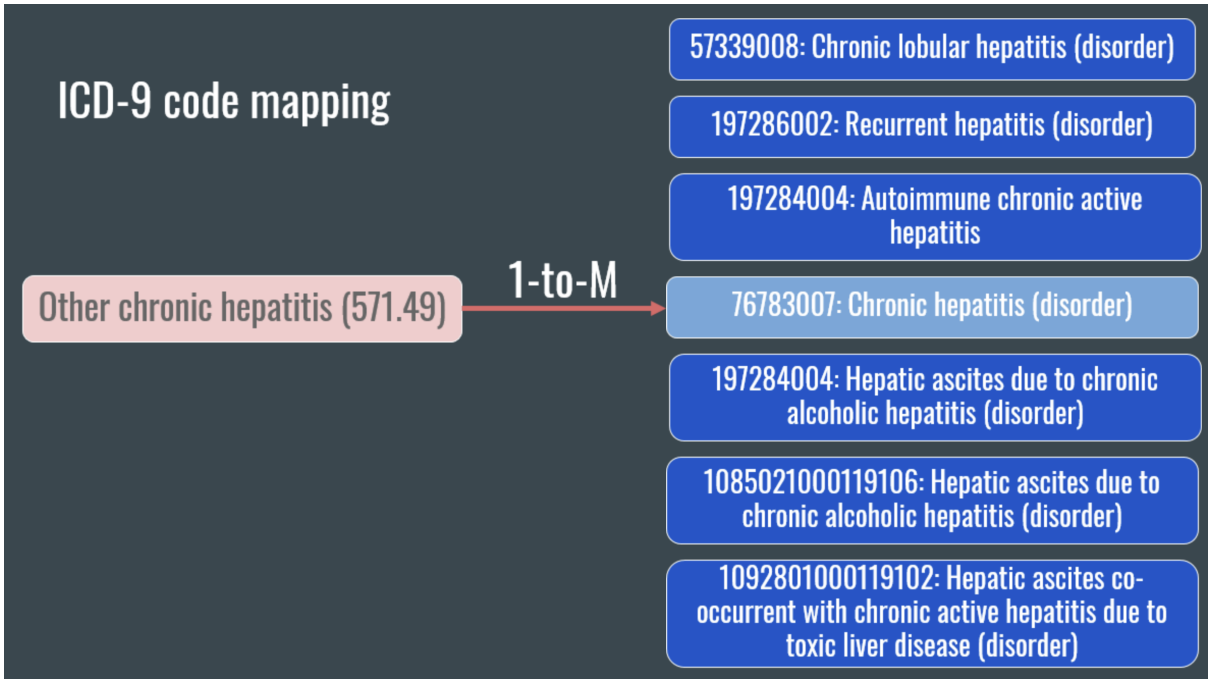

Figure 1: Example on how certain expert ICD-9 translations are mapped to many SNOMED CT codes. Depending on the context, the most accurate SNOMED CT code must be selected.

Each terminology covers different types of medical entities, with different approaches on placing concepts in meaningful hierarchies. While ICD-9 pools detailed diagnoses under more generic ones; NDC constructs medication codes from three other ones for labelers, products, packages. These differences create unique challenges for each dataset, therefore different approaches must be taken to align each one.

3.2. Strategies for Medical Translation

As our training data, whenever an expert-curated 1-to-1 translation to SNOMED CT exists in a public refset, we retrieve it directly. This allows us to leverage golden standards as much as possible. However, the real challenge arises when mappings are noisy, entirely missing, or highly ambiguous. To address this, we developed two separate strategies to handle these scenarios for each terminology we are working with.

3.2.1. Setting A: Predicting Missing Translations for Procedures and Medications

This strategy is devised to handle the vast number of unmapped **CPT** and **NDC** codes. When a code is completely missing from our translation dictionary, we cannot rely on links; we must read the definitions. We utilize a transformer-based AI language model called medBERT [18], which has been fine-tuned on medical texts. Using medBERT, we convert the text descriptions of both the unmapped code (e.g. CPT's "removal of appendix") and potential targets (e.g. SNOMED CT's "appendectomy") into vector embeddings. Essentially plotting the sentences as coordinates in a mathematical space. By measuring the cosine similarity (the angle) between these coordinates, the system understands that

the texts might mean the same thing, even if they use different words. We then feed this similarity score into a trained Logistic Regression model, acting as an automated judge that decides if the pair is a true match. To train this judge, we provided it with examples of correct translations from our refsets (positive pairs) and deliberately mismatched codes (synthetic negative pairs) so it could learn the boundary between a good and a bad translation (see Figure 2).

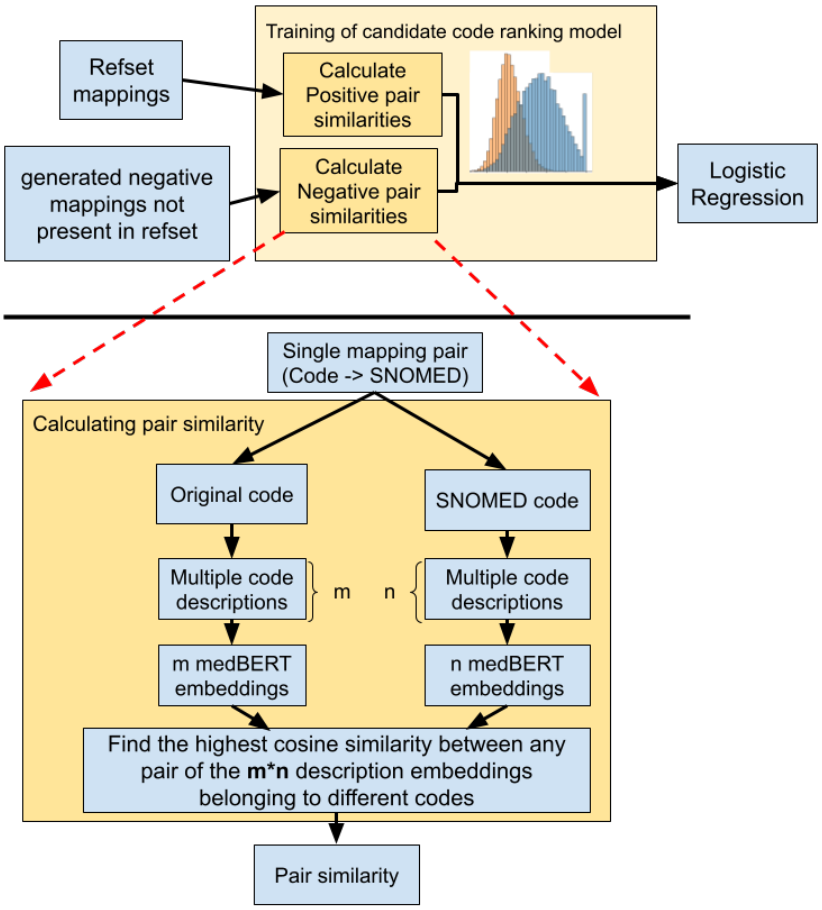

Figure 2: Predicting Missing Translations by Meaning. To find matches for completely unmapped codes (like CPT procedures or NDC drugs), we convert the text descriptions of the codes into mathematical vectors using medBERT. The system finds the highest cosine similarity (meaning overlap) between descriptions. A Logistic Regression model, trained on known good and bad matches, then calculates the probability that the two codes mean the same thing.

3.2.2. Setting B: Enrichment with Patient History for Diagnoses

This strategy solves the ambiguity of **ICD-9** codes. Often, an ICD-9 code maps to many potential SNOMED CT codes (a 1-to-M mapping). A generic ICD-9 code for ”viral infection” might map to codes for ”viral eye infection”, ”viral pneumonia”, or ”viral skin rash” in the target terminology. To pick the right one, we look at the patient’s electronic health record, i.e. their clinical context. For instance, if the patient’s recent history contains codes for ”antiviral eye drops” and ”vision blurriness”, the context

suggests the correct translation should be "viral eye infection". To properly include patient history, we need to measure how closely related the candidate codes are to the patient's context codes within the SNOMED CT network graph. We considered three ways to measure this closeness:

1. **Textual similarity (medBERT):** Found to be superficial, as it ignores the complex medical relationships built into the SNOMED CT network.
2. **Shortest path:** Counting the exact steps between concepts in the SNOMED CT graph. While accurate, calculating paths across 470,000 interconnected nodes was computationally taxing.
3. **Graph Embeddings:** Specifically a node embedding, Node2Vec [19] is a technique adapted from Word2Vec [20]. It explores the SNOMED CT network using random walks and generates spatial coordinates (embeddings) for every concept. This allows us to instantly calculate the semantic similarity (cosine) between any two medical concepts relatively quickly.

Using Node2Vec distances, we execute the context-ranking process (Figure 3). First, we compute distances $d(K_i, C_j)$ between each candidate code $K_i \in \{K_1, K_2, \dots, K_n\}$ and context code $C_j \in \{C_1, C_2, \dots, C_m\}$ via cosine similarity. We collect these into a vector $\mathbf{d}_j = [d(K_1, C_j), d(K_2, C_j), \dots, d(K_n, C_j)]$ to characterize our context. Because raw distance numbers can be skewed, we convert them into an comparable percentile rank ($r_{i,j}$). We then derive a normalized fit score $s_{i,j}$ (ranging from 0 to 1), where 1 means the candidate is a perfect contextual fit (see Equation 1).

$$s_{i,j} = 1 - \frac{r_{i,j} - 1}{n - 1} \tag{1}$$

However, not all patient context is helpful. If a patient had a broken arm ten years ago, that context code is irrelevant to their current "viral infection". To ensure the model ignores irrelevant background noise, we assign an importance weight (w_j) to each context code using the Interquartile Range (IQR) of the distances (see Equation 2 and 3). If a context code is completely unrelated to the current problem, all candidates will be equally distant from it, resulting in an IQR of zero. These nodes are excluded to minimize noise in the context vector.

$$IQR_j = Q_3(\mathbf{d}_j) - Q_1(\mathbf{d}_j) \text{ where } Q_3 \text{ is the third and } Q_1 \text{ is the first quartile} \tag{2}$$

$$w_j = \max(\text{IQR}_j, 0) \tag{3}$$

Finally, we calculate the global ranking score S_i for candidate K_i by taking the weighted average of its fit scores across all helpful context codes (see Equation 4). This guarantees that our final choice is heavily influenced by the patient's relevant medical history, while ignoring distant past events.

$$S_i = \frac{\sum_{j=1}^{m} w_j s_{i,j}}{\sum_{j=1}^{m} w_j + \epsilon} \text{ where } \epsilon = 10^{-6} \text{ for avoiding divisions by zero} \tag{4}$$

3.3. Confidence Calculation

As our translations will not have a ground truth associated with them, we evaluate our results by calculating confidence scores based on the golden standard available to us. We assign a confidence score to every aligned code based on how it was generated:

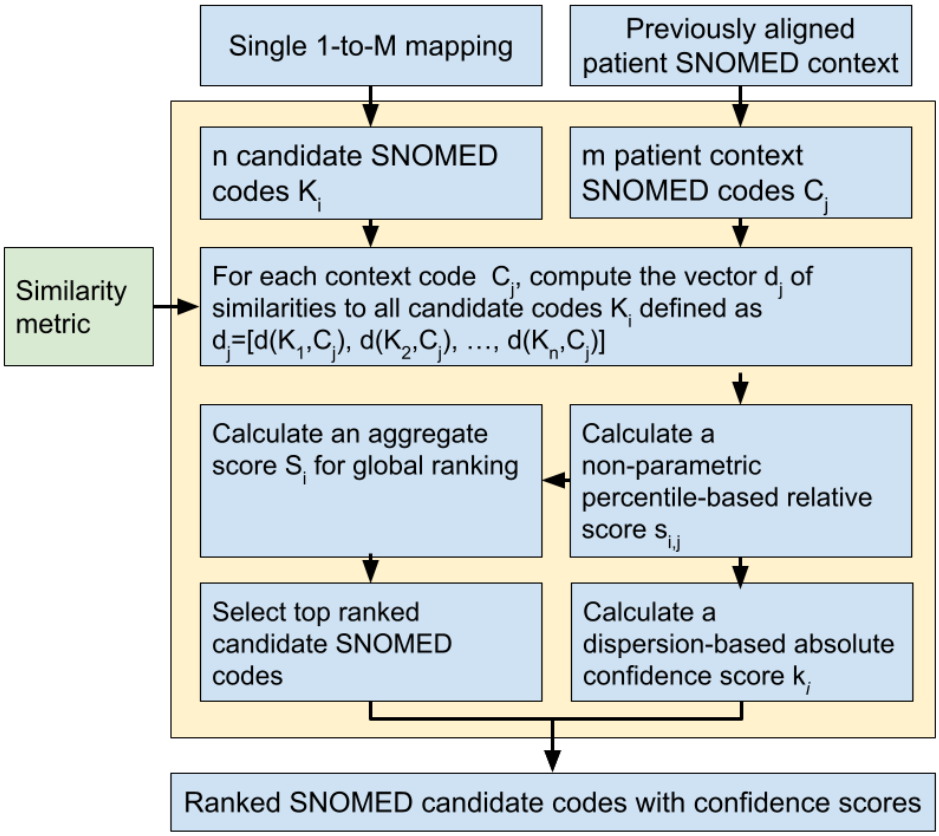

Figure 3: Ranking Ambiguous Codes Using Patient Context. When an ICD-9 code translates into many potential SNOMED CT candidates, we utilize the patient's existing medical history (context codes). We use Node2Vec graph embeddings to instantly measure the mathematical distance between candidate codes and context codes in the SNOMED CT network. By heavily weighting relevant context and ignoring irrelevant noise, we generate an aggregate score to rank the most likely correct translation.

1. **Expert Curated (All 1-to-1 Refsets):** Assigned 100% confidence, as they represent the verified ground truth.
2. **Imputed CPT and NDC Codes:** The confidence score is the probability percentage output directly by the Logistic Regression, i.e. the "judge" model.
3. **Context-Ranked ICD-9 Codes:** Because this ranking depends on a patient's history, we propose a separate confidence metric ($\kappa_i \in [0, 1]$). Our aim is to reward a candidate that is an absolute, undeniable match in at least one highly relevant context, rather than a candidate that is just "mildly okay" across the board. First, we calculate how far a candidate rises above the median average (its positive deviation, $\delta_{i,j}$) as seen in Equation 5.

$$\delta_{i,j} = \max(s_{i,j} - 0.5, 0) \tag{5}$$

We then scale this based on the weight of the context to find our final confidence score (κ_i) (see Equation 6).

$$\kappa_i = 2 \cdot \max_{j=1,\ldots,m} \left(\delta_{i,j} \cdot w_j\right) \tag{6}$$

In plain terms, if a candidate perfectly matches a highly important piece of the patient's medical history (an outlier behavior), it receives a confidence score near 100%. If a candidate never stands out from the rest, its score drops to 0%. This guarantees we only trust the suggestion when it finds a decisive, context-specific winner.

3.4. Experimental Setup

With our translation strategies and confidence metrics mathematically defined, we must rigorously validate how well these methods perform in their respective experimental settings.

Experiment A: Testing the Judge for Procedures and Medications. To evaluate how well the Logistic Regression model from Strategy B can guess missing translations for procedure (CPT) and medication (NDC) codes, we analyze its discriminative power. Specifically, we train four distinct variations of the model to see which learns the boundaries of medical meaning best. We compare models trained exclusively on perfect, 1-to-1 expert translations against others trained on a mix of 1-to-1 and messier 1-to-M translations. By testing these variations on both CPT and NDC datasets, we can indirectly measure how accurately the model distinguishes a true medical synonym from a randomly generated false one.

Experiment B: Tuning the Contextual Translations for Diagnoses. For Setting B, our success relies entirely on the Node2Vec graph embeddings' accuracy. To ensure our cosine similarity accurately reflects true medical relationships, we must fine-tune its hyperparameters. We do this by calculating the Spearman correlation [21] (ρ) between the shortest path between two nodes in the SNOMED CT hierarchy and our derived similarity score. As a shorter walking distance (closely related concepts like "viral pneumonia" and "lung infection") should naturally result in a higher similarity score, an ideal ρ would be close to -1. Due to the immense computational cost of calculating exact paths across the entire 470,000-node SNOMED CT terminology, we run these tuning experiments on a dense, representative sub-hierarchy of 3,000 highly connected medical concepts. This smaller testing ground allows us to rapidly find the optimal settings before deploying the map globally. In other words, this experiment allows us to find the best state of our method for the entity alignment task.

4. Results

Having established the entity alignment framework and its experimental setup, the following sections illustrates the successes and limitations of Settings A and B.

Experiment A We first analyzed the discriminative power of our Logistic Regression approach for aligning procedure and medication codes. The model tasked with guessing missing translations for procedures (CPT) and medications (NDC) based solely on their text descriptions. This highlighted how reliably the model could separate true, expert-verified matches (positive pairs) from randomly generated, incorrect matches (negative pairs).

The results showed a stark contrast between CPT and NDC codes. For medications, the textual embeddings were highly effective. As seen in Figure 4, the model clearly distinguished the true mappings from the false ones, resulting in very little overlap between the two categories. This suggests that medications' descriptions are semantically distinct enough for text-based vectors to align them accurately.

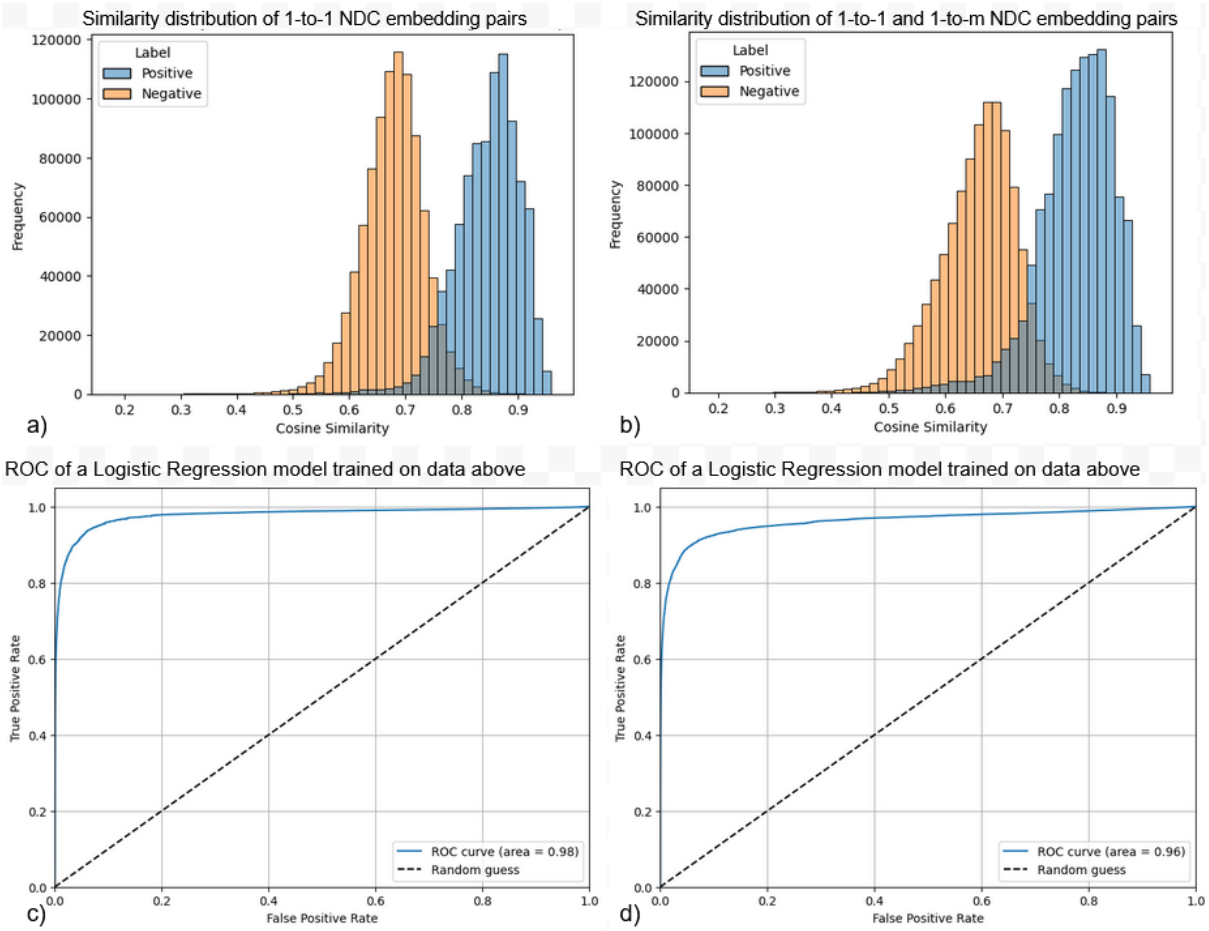

Figure 4: Performance on Drug Translations (NDC). The histograms show how well the model separates true translations (blue) from false ones (orange). The minimal overlap demonstrates high accuracy, indicating that medications' descriptions are highly distinct. Subfigure (a) shows the model used for 1-to-M ranking, while (b) shows the model used for imputing missing links.

Conversely, procedures proved far more challenging. Figure 5 reveals a significant overlap between true and false matches. This blending occurs because many different medical operations share identical or highly similar textual descriptions in the data. When the text is identical but the underlying medical meaning is slightly different, our approach which relies solely on textual descriptions struggles to pick the definitive 1-to-1 match.

Experiment B Next, we evaluated our strategy of using patients' medical histories as context to resolve highly ambiguous diagnosis codes presented in ICD-9 terminology. Before deploying our

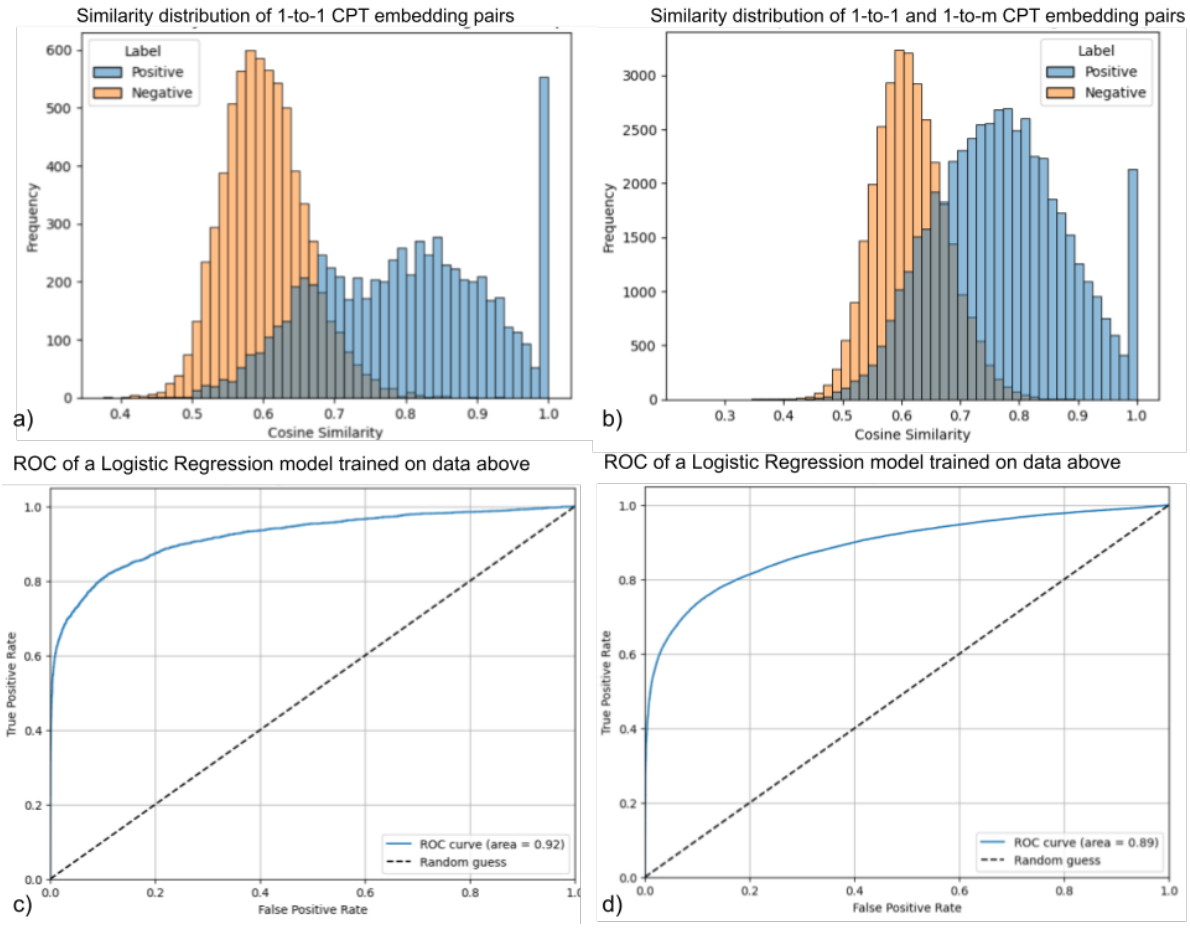

Figure 5: Performance on Procedure Translations (CPT). Procedure descriptions frequently overlap in text. This causes true translations (blue) and false translations (orange) to mix, making it harder for the text-based model to distinguish between them. Subfigure (a) shows the model used for 1-to-M ranking, while (b) shows the model used for imputing missing links.

Node2Vec graph embeddings, we determined its optimal settings using the Spearman correlation metric. Figure 6 illustrates the tuning results on a 3,000-node test set. By comparing our scores against shortest paths between two nodes in the SNOMED CT hierarchy, we found that configuring the model to take 10 random walks of length 10 per node yielded the best correlation. We then applied these optimal settings to align the entire 470,000-node SNOMED CT ontology to the ICD-9 taxonomy.

With our Node2Vec embeddings calibrated, we compared how well three different similarity metrics could rank the correct diagnosis using the patient's context. The results, shown in Table 1, demonstrate the value of structural context. Both metrics that aligned the actual SNOMED CT hierarchy relationships vastly outperformed the method that relied solely on text descriptions, i.e. using medBERT. Most importantly, Node2Vec aligned the high accuracy of the Shortest Path calculation after just one training epoch, while offering a massive computational advantage: Node2Vec can process 1,000 patient rankings in just a second, whereas calculating exact shortest paths takes eight minutes to complete the exact same task.

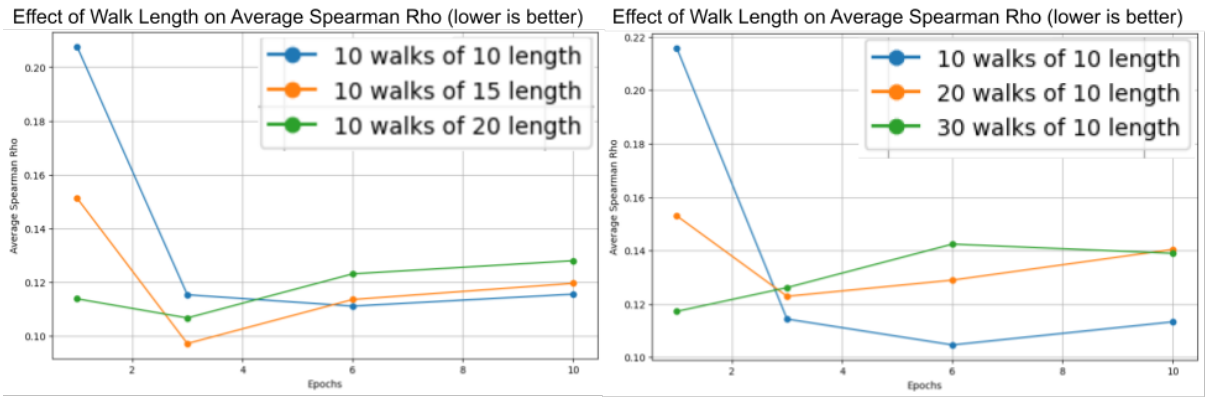

Figure 6: Tuning Node2Vec Embeddings. This chart displays the error rate (Spearman correlation ρ) across different Node2Vec parameter settings on a small subset of the network. Lower values indicate an alignment that better reflects true distances. Ten walks of length ten provided the optimal balance and were chosen for the final model.

		Similarity Metric and Scores		
Test Set	Performance Metric	medBERT Text Embeddings	Shortest SNOMED CT Graph Distances	Node2Vec Graph Embeddings
Set 1	Hits@1	0.06 - 0.12	0.44 - 0.52	0.37 - 0.45
Set 1	Hits@5	0.53 - 0.63	0.82 - 0.88	0.82 - 0.88
Set 2	Hits@1	0.21 - 0.27	0.29 - 0.35	0.28 - 0.36
Set 2	Hits@5	0.69 - 0.75	0.75 - 0.79	0.75 - 0.81

Table 1
Comparing Translation Methods Using Patient Context. This table shows the accuracy of matching ambiguous diagnoses. Graph-based methods (Shortest Distances and Node2Vec) vastly outperform text-only methods (medBERT). Note that Set 1 and Set 2 refer to specific experimental partitions (scores shown ± 1 standard deviation across 10 repetitions).

5. Discussion

The results reflect our choice of using multiple methods for tackling unique challenges of different medical terminologies. Relying solely on expert-curated refsets is insufficient due to missing entries, while relying entirely on text-matching ignores the rich, structured knowledge built into medical networks.

CPT vs. NDC. The performance of our Logistic Regression model highlighted a fascinating difference between translating procedures (CPT) and medications (NDC). Our approach easily separated true medication translations from false ones as pharmaceutical descriptions are highly diverse and chemically specific. Conversely, medical procedures frequently use generic, overlapping vocabulary (e.g. ”excision of lesion”), which artificially inflates their text similarity and confuses embeddings. This tells us that using only the textual context can be sufficient for aligning medication codes, but less reliable for

procedures due to limited and similar contexts.

Leveraging Tuned Node2Vec Embeddings. When we attempt to translate highly ambiguous diagnosis codes (ICD-9), looking at the patient's context proved to be crucial. As Table 1 presents, relying solely on textual context using medBERT to characterize patient history performs poorly. However, when we equip the system with a mathematical alignment of the SNOMED CT network, using either the exact Shortest Paths between nodes or our Node2Vec approximation, accuracy significantly increases. This highlights that understanding the structural "is-a" and "caused-by" relationships between medical concepts is far more valuable for matching diagnoses than simply reading their textual descriptions. Our approach using Node2Vec successfully captures this structural knowledge while remaining fast enough for real-world application.

5.1. Limitations

Although our algorithmic two-method framework for aligning medical terminologies yields interesting results, several major limitations must be noted. Firstly, to evaluate our context-ranking strategy, we relied on curated patient histories from MIMIC-III. While this provided a controlled testing ground, real-world electronic health records are infamously messy and complex, which may impact the model's accuracy in a live hospital setting. Third, our alignment framework fundamentally relies on having a starting point, i.e. it requires some existing expert dictionaries (refsets) to know which neighborhood of the map to look in. It may struggle to adapt to entirely new, emerging medical terminologies that have no prior expert curation. Finally, due to severe computational limits, our Node2Vec map was only trained for a single pass (one epoch) over the massive 470,000-node SNOMED CT network, meaning our approach is yet to reach its top performance.

6. Conclusions & Future Work

In this study, we have presented a two-method framework for medical entity alignment. By combining expert dictionaries with algorithmic and statistical approaches, we manage to align fragmented medical languages into the unified SNOMED CT master terminology. We cover procedures (CPT), medications (NDC), and diagnoses (ICD-9) which required different methods for addressing challenges unique to each terminology.

We propose two methods. First, we demonstrated that a medBERT-powered Logistic Regression model can efficiently guess and rank missing translations for unmapped drugs and procedures. Even when textual descriptions overlap heavily, this approach learns the boundaries of medical vocabulary to find the correct match. Finally, to tackle highly ambiguous diagnosis codes, we implemented a context-aware ranking system using Node2Vec embeddings. We proved that mapping the structural "is-a" and "caused-by" relationships of medical concepts vastly outperforms simply reading their text descriptions. Our alignments matched the accuracy of exhaustive, exact-path calculations, achieving Hits@5 scores of 0.82-0.88 while operating with a low computational cost. Moreover, our framework does not just make provide alignment suggestions from SNOMED CT; it tells us how much to trust those guesses. By calculating a transparent confidence score (κ_i), we explicitly highlight when a translation is a robust, undeniable match based on a patient's unique medical history.

Moving forward, our immediate goal is allowing the Node2Vec method from Setting B to train for a full 10 epochs. We estimate this deeper training will take approximately 75 hours on a standard

processor, but it is expected to yield a significantly sharper and more accurate map from ICD-9 to SNOMED CT. Furthermore, instead of using Node2Vec which treats the input as undirected graphs, we can upgrade to RDF2Vec[9] which works on directed graphs to improve our overall accuracy. Finally, we plan to introduce configurable precision-recall thresholds derived from the ROC curves for our imputation models from Setting A. This will allow clinicians to tailor the system to their specific needs, balancing whether they want to safely capture all possible code translations or show only the most specific, absolutely certain matches.

Ultimately, by bridging the gaps left by incomplete expert dictionaries, our framework allows computer systems to capture patient context across different hospital departments. We have validated that capturing the medical terminologies' hierarchies is essential for accurate translation. For the broader healthcare community, these advancements promise significantly fewer coding errors, cleaner datasets, and dramatically faster analysis of electronic health records, paving the way for smarter, more connected patient care.

7. Declaration on Generative AI

Generative AI tools were used to draft several sections of the manuscript which were significantly edited before the final draft.

Acknowledgments

This work is supported by the Horizon Europe Framework Program under Grant Agreement No. 101057062 (AIDAVA). SNOMED CT licenses are handled through this project.

References

[1] R. Sahay, D. Ntalaperas, E. Kamateri, P. Hasapis, O. D. Beyan, M.-P. F. Strippoli, C. A. Demetriou, T. Gklarou-Stavropoulou, M. Brochhausen, K. Tarabanis, T. Bouras, D. Tian, A. Aristodimoux, A. Antoniadesx, C. Georgousopoulos, M. Hauswirth, S. Decker, An ontology for clinical trial data integration, in: 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 3244–3250. doi:10.1109/SMC.2013.553.

[2] S. Mate, F. Köpcke, D. Toddenroth, M. Martin, H.-U. Prokosch, T. Bürkle, T. Ganslandt, Ontology-based data integration between clinical and research systems, PLOS ONE 10 (2015) 1–20. doi:10.1371/journal.pone.0116656.

[3] I. de Zegher, K. Norak, D. Steiger, H. Müller, D. Kalra, B. Scheenstra, I. Cina, S. Schulz, K. Uma, P. Kalendralis, et al., Artificial intelligence based data curation: enabling a patient-centric european health data space, Frontiers in medicine 11 (2024) 1365501.

[4] R. Ambalavanan, R. S. Snead, J. Marczika, G. Towett, A. Malioukis, M. Mbogori-Kairichi, Ontologies as the semantic bridge between artificial intelligence and healthcare, Frontiers in Digital Health 7 (2025) 1668385.

[9] http://www.rdf2vec.org/

[5] V. Dimitrieski, G. Petrović, A. Kovačević, I. Luković, H. Fujita, A survey on ontologies and ontology alignment approaches in healthcare, in: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, 2016, pp. 373–385.

[6] E. Jiménez-Ruiz, B. Cuenca Grau, Logmap: Logic-based and scalable ontology matching, in: L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, E. Blomqvist (Eds.), The Semantic Web – ISWC 2011, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 273–288.

[7] M. Chen, Y. Tian, M. Yang, C. Zaniolo, Multilingual knowledge graph embeddings for cross-lingual knowledge alignment, arXiv [cs.AI] (2016). doi:10.48550/ARXIV.1611.03954.

[8] Z. Sun, W. Hu, C. Li, Cross-lingual entity alignment via joint attribute-preserving embedding, arXiv [cs.CL] (2017). doi:10.48550/ARXIV.1708.05045.

[9] Z. Sun, W. Hu, Q. Zhang, Y. Qu, Bootstrapping entity alignment with knowledge graph embedding, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, California: International Joint Conferences on Artificial Intelligence Organization (2018). doi:10.24963/ijcai.2018/611.

[10] M. Chen, Y. Tian, K.-W. Chang, S. Skiena, C. Zaniolo, Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment, arXiv [cs.AI] (2018). doi:10.48550/ARXIV.1806.06478.

[11] X. Tang, J. Zhang, B. Chen, Y. Yang, H. Chen, C. Li, Bert-int: A bert-based interaction model for knowledge graph alignment, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, California: International Joint Conferences on Artificial Intelligence Organization (2020). doi:10.24963/ijcai.2020/439.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[13] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, D. Zhao, Relation-aware entity alignment for heterogeneous knowledge graphs, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2019.

[14] X. Mao, W. Wang, H. Xu, Y. Wu, M. Lan, Relational reflection entity alignment, Proceedings of the 29th ACM International Conference on Information & Knowledge Management, New York, NY, USA: ACM (2020). doi:10.1145/3340531.3412001.

[15] Y. Wu, X. Liu, Y. Feng, Z. Wang, D. Zhao, Neighborhood matching network for entity alignment, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics (2020). doi:10.18653/v1/2020.acl-main.578.

[16] Z. Liu, Y. Cao, L. Pan, J. Li, Z. Liu, T.-S. Chua, Exploring and evaluating attributes, values, and structures for entity alignment, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Stroudsburg, PA, USA: Association for Computational Linguistics (2020). doi:10.18653/v1/2020.emnlp-main.515.

[17] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016) 160035.

[18] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ digital medicine 4 (2021) 86.

[19] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, CoRR abs/1607.00653 (2016). URL: http://arxiv.org/abs/1607.00653. arXiv:1607.00653.

[20] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv (2013). URL: https://doi.org/10.48550/arXiv.1301.3781.
[21] J. Zar, Spearman Rank Correlation, volume 5, 2005. doi:10.1002/0470011815.b2a15150.