Original Research

Contrastive Learning for Clinical Sentence Similarity Estimation in Medical Question Answering Systems

Hande Erkoç[1]

[1]Uşak Institute of Technology, Department of Computer Science, İstiklal Mah. 75. Cadde No:16, Uşak, Turkey.

Abstract

This paper explores the application of contrastive learning to estimate clinical sentence similarity in the domain of medical question answering systems. The goal is to improve the accuracy and reliability of automated tools that respond to complex questions posed by healthcare professionals, clinical researchers, and patients. By focusing on sentence-level embeddings within clinical text corpora, our approach emphasizes the subtle linguistic cues and domain-specific contextual factors that determine semantic similarity in medical dialogues. We present a robust framework that leverages a contrastive objective to maximize the alignment between semantically related sentences while preserving essential distinctions among dissimilar examples. Additionally, we incorporate advanced representation learning techniques and rigorous optimization strategies to enhance the encoding of nuanced medical terminology. This paper addresses several core challenges: capturing long-range dependencies in clinical discourse, handling synonyms and abbreviations common to the healthcare domain, and mitigating the impact of noisy electronic health records on model performance. Our results show that a carefully designed contrastive learning pipeline yields significantly higher similarity estimation accuracy than standard sentence embedding baselines, with notable improvements in precision for semantically complex queries. We also provide a theoretical perspective on the relationship between contrastive objectives and the underlying geometry of sentence embeddings. Finally, we discuss the implications of our findings for broader clinical text mining applications.

1. Introduction

The emergence of large-scale medical text repositories, including electronic health records, scientific publications, and patient-generated content, has created a burgeoning need for computational methods that can efficiently parse, interpret, and analyze such materials [1]. Medical question answering systems have consequently evolved into highly specialized tools that can support evidence-based practice, enhance clinical decision-making, and enable self-directed patient education [2]. Central to many of these systems is the core task of sentence similarity estimation, whereby a query, often expressed in colloquial or semi-technical language, must be matched to the most relevant text in a medical knowledge base. However, the clinical domain presents unique semantic and pragmatic challenges: for example, medical texts often feature specialized abbreviations, domain-specific jargon, and regionally variable expressions [3]. These factors complicate the mapping from query sentences to their most appropriate answers, making robust methods for sentence similarity estimation indispensable.

The present work proposes a contrastive learning approach designed to capture intricate semantic relationships among sentences that arise in clinical contexts [4]. By harnessing a contrastive objective, the framework systematically aligns pairs of sentences that convey similar medical concepts while ensuring that irrelevant or dissimilar sentences remain appropriately differentiated in the embedding space. If s_1 and s_2 denote two clinically coherent expressions, the objective seeks to minimize a metric $d(\mathbf{h}(s_1), \mathbf{h}(s_2))$, where $\mathbf{h}$ is the encoding function that maps textual inputs to vectors in $\mathbb{R}^n$. For any pair of unrelated sentences s_3 and s_4, the objective likewise maximizes $d(\mathbf{h}(s_3), \mathbf{h}(s_4))$ to maintain a clear semantic boundary. In medical discourse, seemingly small lexical or syntactic differences can drastically alter meaning (for instance, distinguishing “Type I diabetes” from “Type II diabetes”) [5]. Thus, the geometry of the embedding space must be keenly sensitive to these distinctions, yet tolerant of equivalent or near-equivalent phrasings that hinge on variations in clinical syntax.

The utility of sentence similarity estimation in medical question answering systems is multi-fold [6]. It can facilitate more responsive and context-aware query expansions, refine document retrieval pipelines, and yield improved precision in identifying authoritative medical content for end users. Robust sentence similarity methods can also expedite the creation of clinical knowledge graphs [7], which demand consistent and reliable mappings between semantically related entities [8]. For instance, let Q_i represent a query about medication side effects and C_j a corpus sentence describing drug contraindications. A similarity function $\text{Sim}(Q_i, C_j)$ can serve as the backbone of a retrieval system that directs a clinician to information on adverse events and recommended courses of action.

Despite significant strides in natural language processing, there remain numerous domain-specific hurdles [9]. Clinical texts often contain unstructured data peppered with typographical errors and partial forms of relevant medical terms. The variability in abbreviations, such as “BP” for blood pressure versus “BP” for bipolar disorder, necessitates a representation technique that adeptly handles polysemy [10]. The introduction of advanced large language models has been transformative in many general domains but can yield unpredictable results in specialized domains like healthcare if not properly adapted. Domain adaptation typically requires substantial in-domain pretraining on medical corpora, such that the model’s parameters capture the nuance of medical language [11]. The complexity is further compounded by privacy constraints that often limit the dissemination of patient records needed for training high-capacity models.

In light of these challenges, the contrastive learning paradigm offers a promising avenue [12]. When provided with pairs of sentences known to be semantically close (for instance, a medical question and its corresponding expert answer), a contrastive learner updates its parameters to bring these sentences closer in the embedding space. Conversely, unrelated or misleading pairs are pushed farther apart [13]. Formally, if we define a set of labeled pairs $\{(s_i^+, s_i^-)\}$, where s_i^+ are semantically matching sentences and s_i^- are randomly sampled non-matching sentences, the training objective is typically expressed as a margin-based or log-based contrastive loss. A margin-based loss might take the form

$$\sum_i \max\bigl(0, \alpha + d(\mathbf{h}(s_i^+), \mathbf{h}(s_i^-)) - d(\mathbf{h}(s_i^+), \mathbf{h}(s_i^+))\bigr),$$

where α denotes a pre-specified margin and $d(\cdot, \cdot)$ is a distance function, often the Euclidean or cosine distance in a high-dimensional embedding space [14]. Such loss functions enforce a separation between positive and negative pairs. Log-based forms like the InfoNCE loss further incorporate normalization terms that facilitate gradient-based optimization in neural models. [15]

Another pivotal consideration is model architecture [16]. Transformer-based encoders have emerged as a de facto choice for numerous language processing tasks, owing to their self-attention mechanism that effectively captures long-range dependencies. However, in the realm of medical text, these architectures can be prone to overfitting if pretrained only on general corpora such as Wikipedia or Common Crawl [17]. Fine-tuning on domain-specific corpora like PubMed abstracts or clinical notes from specialized repositories often mitigates this mismatch. Let $\mathbf{E} \in \mathbb{R}^{d \times n}$ represent the token embeddings of a clinical sentence, where d is the dimensionality of each token embedding and n is the sentence length. The Transformer computes self-attention weights via [18]

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Bigl(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\Bigr)\mathbf{V},$$

where $\mathbf{Q} = \mathbf{W}_Q\mathbf{E}$, $\mathbf{K} = \mathbf{W}_K\mathbf{E}$, and $\mathbf{V} = \mathbf{W}_V\mathbf{E}$ are learned parameter matrices. This mechanism refines token-level interactions, capturing cross-dependencies such as medical abbreviations resolving to expanded forms, or context cues that shift the interpretation of a particular medical condition.

The remainder of this paper is structured to delve deeper into these considerations, culminating in a detailed analysis of how contrastive learning for clinical sentence similarity can be both empirically robust and theoretically justifiable [19]. In the subsequent sections, we outline the foundational principles of medical language understanding, introduce our proposed contrastive learning approach, present our experimental results on a variety of real-world clinical corpora, and finally analyze the theoretical underpinnings of sentence encoding within this domain. The overarching objective is to move toward a more nuanced and faithful computational representation of medical discourse, thereby advancing the capabilities of automated question answering in healthcare contexts. [20]

2. Foundations of Medical Language Understanding

Technical progress in medical language understanding has been propelled by innovations in both linguistic representations and neural network architectures. A key driver of these advances has been the recognition that natural language in a clinical setting follows complex contextual and structural norms [21]. These norms guide how diagnoses, treatments, and outcomes are reported, aggregated, or cross-referenced. To navigate these norms effectively, models must handle phenomenon such as abbreviation resolution, domain-specific entity recognition, and the interplay between domain-agnostic linguistic features and domain-centric conceptual frameworks. [22, 23]

The concept of layered linguistic representation often underpins modern approaches. If we let x denote a clinical sentence, it can be decomposed into tokens $x = (w_1, w_2, \dots, w_n)$ [24]. Each token w_i may correspond to a medical entity (e.g., a drug name) or a descriptive term (e.g., an adjective describing a symptom). Embeddings typically map w_i to a vector $\mathbf{e}_i \in \mathbb{R}^d$. Simple word embeddings might treat $\mathbf{e}_i$ as a static lookup from a large vocabulary. However, context-sensitive embeddings use surrounding words to refine each $\mathbf{e}_i$. For example, consider a sentence describing a patient's "BP," which might stand for "blood pressure" or "bipolar disorder" depending on context [25]. Let $C(w_i)$ represent the set of context words for w_i. A context-aware embedding function $\mathbf{h}(w_i, C(w_i))$ must assign distinct internal representations based on how "BP" is used in the sentence.

Efforts to capture domain-specific knowledge have included specialized embeddings such as BioWordVec, SciBERT, and ClinicalBERT, which incorporate text from biomedical papers and clinical notes [26]. Such models attempt to ground tokens in both linguistic and medical ontological frameworks. For instance, "diabetic ketoacidosis" might be recognized as a term highly correlated with insulin deficiency and blood glucose measurements [27]. Let $\mathbf{v}$ be the embedding vector for "diabetic ketoacidosis," and let $\mathbf{r}$ be an embedding vector for an insulin-related concept. A measure of similarity $\langle \mathbf{v}, \mathbf{r} \rangle$ might indicate the conceptual closeness of these medical expressions, informing downstream tasks such as question answering where user queries about diabetic emergencies should retrieve relevant passages containing "diabetic ketoacidosis."

From a theoretical standpoint, the representation space can be viewed as a manifold where semantically related tokens or sentences cluster. In the broader context of question answering, the objective is to retrieve the correct cluster or region of the manifold that best matches a given query [28]. Logical formalisms often describe this process in terms of inference rules. Let ϕ be a logical formula denoting a set of clinical assertions, and let ψ be a user query that must be resolved against these assertions [29]. One might say $\phi \models \psi$ if ψ logically follows from ϕ. Translating this into an embedding-based paradigm, the question becomes whether the vector representations of ϕ and ψ align to a degree exceeding a predetermined threshold. [30]

The question of how to optimize these embeddings remains central [31]. Gradient-based methods, particularly those derived from backpropagation through large neural architectures, have shown substantial performance gains. However, large-scale training on clinical corpora is often constrained by data availability, confidentiality requirements, and the risk of overfitting if the corpus is not sufficiently diverse [32]. Consequently, data augmentation and semi-supervised strategies have become popular, whereby unlabeled clinical texts are used to refine representations via self-supervised tasks. These tasks might include masked language modeling, in which a fraction of tokens is randomly masked and the model is trained to predict them, or next-sentence prediction, which encourages an understanding of sentence-level coherence. [33]

A challenge that emerges in domain adaptation is how to handle distributional shifts. The textual data found in clinical settings can be drastically different from that in open-domain corpora [34]. Formal logic statements sometimes help clarify the constraints: let D_g be the general domain data distribution and D_c be the clinical domain distribution. We consider a statement: $\forall x \in X, P(x \mid D_c) \neq P(x \mid D_g)$ [35]. This essentially indicates that probabilities of textual sequences in the clinical domain can differ significantly from their general-domain counterparts. Models pretrained on D_g alone can fail to capture these differences, leading to inaccurate embeddings and suboptimal question answering performance [36]. Consequently, bridging the gap involves fine-tuning on labeled or unlabeled data from D_c such that the learned representation function $\mathbf{h}_c$ respects the unique constraints and style of clinical text.

Moreover, the intricacy of medical language understanding extends to capturing phenomena such as negation. Sentences that differ only by a negation term can map to starkly different semantic interpretations [37]. For example, “The patient does not have pneumonia” and “The patient has pneumonia” should be mapped to distinct regions in the embedding manifold. This requirement often leads to specialized modules or training protocols that highlight negation detection and representation [38]. Another dimension is the handling of numeric values and lab results (e.g., blood glucose levels). A slight difference in numeric terms can change the entire interpretation of a sentence [39]. Models must be sensitive to these changes, possibly by maintaining specialized embeddings or gating mechanisms that handle numeric terms.

The complexities intrinsic to medical language understanding underscore the need for a robust and flexible framework that can adapt to diverse syntactic, semantic, and pragmatic features [40]. Contrastive learning, the focus of this paper, is one such framework. By imposing a structured learning objective that rewards the closeness of semantically aligned sentences and penalizes the proximity of unrelated ones, contrastive methods can refine the embedding space in a way that is well-suited to the domain’s complexities [41]. This approach thus sets the stage for enhancing sentence similarity estimation within medical question answering systems.

3. Contrastive Learning Approaches for Sentence Similarity

Contrastive learning methods revolve around building representations where positive examples are mapped closely, and negative examples are pushed apart [42]. At its core lies the concept that semantic similarity must be well-reflected in the geometry of the embedding space. By employing pairs (or triplets) of sentences, the training algorithm iterates through examples where lexical and semantic similarities are known, incrementally shaping the embedding model to respect these relationships in a consistent manner. [43]

Let (x_i, y_i) be a pair of clinically similar sentences, and let (x_i, z_i) be a pair of clinically dissimilar sentences [44]. We define a contrastive loss $\mathcal{L}$ that draws together the representations of x_i and y_i while repelling x_i and z_i. A typical form of the loss function includes a term such as

$$\log \frac{\exp(\mathrm{sim}(\mathbf{h}(x_i), \mathbf{h}(y_i)))}{\exp(\mathrm{sim}(\mathbf{h}(x_i), \mathbf{h}(y_i))) + \exp(\mathrm{sim}(\mathbf{h}(x_i), \mathbf{h}(z_i)))}.$$

Here, $\mathrm{sim}(\cdot, \cdot)$ might be a dot product or cosine similarity. Such a log-softmax expression encourages the model to assign a higher similarity to positive pairs than to negative pairs [45]. Variations of this approach include temperature scaling, which adjusts the “sharpness” of the output distribution, or multiple negative sampling strategies that use large sets of negative examples.

In the context of medical question answering, collecting positive sentence pairs often involves leveraging expert-annotated datasets [46]. For instance, questions and their official reference answers can form the positive pairs, while random sentences from the corpus can function as negative examples. Alternatively, some negative examples can be sampled from domain-specific confounders, ensuring that they are not trivially unrelated but pose a legitimate semantic challenge [47]. Suppose (q_i, a_i) is a question-answer pair deemed correct, and a_j is an answer to another question but shares lexical overlap with a_i. Then (q_i, a_j) might serve as a hard negative pair, forcing the model to differentiate subtle domain shifts. [48]

These approaches often rely on specialized encoders, frequently adapted from large pretrained models such as BERT, to generate embeddings for sentences. A forward pass through such a model transforms a medical sentence s into a high-dimensional representation $\mathbf{h}(s)$. Contrastive learning iteratively refines this encoder based on the proximity or distance of relevant pairs [49]. From a functional perspective, consider that $\mathbf{h} : \mathcal{S} \to \mathbb{R}^d$. The training aims to ensure that, for all relevant pairs (s_1, s_2), the induced metric $d(\mathbf{h}(s_1), \mathbf{h}(s_2))$ is small if they express the same concept, and large otherwise. Formally, one might define a rule:

$$\forall s_1, s_2 \in \mathcal{S}, \quad \begin{cases} \text{SemanticMatch}(s_1, s_2) \Rightarrow d\big(\mathbf{h}(s_1), \mathbf{h}(s_2)\big) < \gamma_1, \\ \neg\, \text{SemanticMatch}(s_1, s_2) \Rightarrow d\big(\mathbf{h}(s_1), \mathbf{h}(s_2)\big) > \gamma_2. \end{cases}$$

where γ_1 and γ_2 are thresholds controlling how close or distant the pairs should be. [50]

Once trained, such contrastive encoders can be deployed in question answering pipelines to quickly locate relevant sections of large clinical corpora. Given a new question q, we compute $\mathbf{h}(q)$, then retrieve candidate answers whose embeddings $\mathbf{h}(a)$ exhibit a high similarity with $\mathbf{h}(q)$. By narrowing the search space, the system can achieve rapid and accurate retrieval, which is critical in time-sensitive environments like clinical decision support [51]. Additional ranking layers or verification strategies may then refine these candidate answers based on domain-specific heuristics, but the initial embedding-based retrieval is often a crucial bottleneck.

Some advanced variants of contrastive learning address data scarcity or domain mismatch challenges by integrating data augmentation or multi-task objectives [52]. For example, Mixup techniques might mix embeddings of two sentences to generate “interpolated” training examples. Though rarely explored in general text domains, this approach can be especially beneficial in medical text due to the high cost of labeled data [53]. Another strategy is to blend contrastive objectives with a masked language modeling objective, thereby balancing the benefits of local token-level supervision and global sentence-level supervision. Let $\mathcal{L}_{\text{mask}}$ denote the loss from masked language modeling, and let $\mathcal{L}_{\text{cont}}$ denote the contrastive loss. A multi-objective approach might optimize [54]

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{cont}} + (1 - \lambda)\mathcal{L}_{\text{mask}},$$

where λ is a scalar hyperparameter. This not only aligns sentences at the embedding level but also ensures robust token representations that generalize well to various downstream tasks. [55]

The success of contrastive learning also depends on hyperparameter tuning, such as the size of the mini-batch, the choice of negative sampling strategy, and the dimensionality of the final sentence representation. Medical text, often replete with domain-specific expressions, might benefit from large mini-batches containing diverse examples of medical terminology [56]. Meanwhile, carefully curated negative examples help the model learn fine-grained distinctions. For instance, the phrase “Type I diabetes” and “Type II diabetes” share a large lexical overlap but differ in pathophysiology [57]. Handling such near-duplicate pairs in the training data ensures the encoder remains discriminative even among closely related medical terms. [58]

In summary, contrastive learning stands out for its ability to produce sentence representations that capture domain-sensitive nuances with minimal reliance on explicit labeling strategies. Its versatility, when integrated with the complexities of medical text, paves the way for more accurate sentence similarity estimation in question answering systems [59]. The method’s success draws from meticulous selection of positive and negative pairs, strategic model architecture design, and appropriate data augmentation practices. The following sections will detail how this approach is operationalized in real-world clinical datasets, evaluate its performance against robust baselines, and offer deeper insights into the theoretical aspects of embedding-based similarity estimation. [60]

4. Experimental Setup and Results

The empirical evaluation aimed to measure the efficacy of our contrastive learning framework in accurately predicting semantic similarity of clinical sentences for question answering tasks. We assembled a multi-faceted dataset comprising pairs of questions and reference answers, each annotated by medical experts [61]. These annotated pairs provided ground-truth labels indicative of the sentences' semantic alignments. The dataset also included negative pairs, carefully curated to ensure a range of lexical overlaps, thus posing a substantive challenge to naive methods. [62]

We selected two broad categories of question-answer pairs. The first category, labeled "Diagnostic Queries," contained questions about symptoms, possible diagnoses, and suggested tests [63]. Typical queries in this category took forms like, "What is the recommended evaluation for a patient presenting with chest pain?" The second category, labeled "Treatment Queries," encompassed questions on recommended drugs, dosing guidelines, and outcome predictions for various medical conditions. These included questions such as, "Which medications are indicated for acute bronchitis?" or "What is the optimal dosage of drug X in pediatric cases?" Both categories spanned a wide range of synonyms, abbreviations, and polyequivalences found in medical texts. [64]

The data preparation phase involved splitting the corpus into training, validation, and test sets. For each question q, we designated an official correct answer a^+ to form a positive pair [65]. We randomly sampled an incorrect or contextually distant answer a^- to form a negative pair. Hard negative examples were included by selecting answers that overlapped lexically with a^+ but addressed a different medical question [66]. Let $\mathrm{LexOverlap}(a^+, a^-)$ denote the ratio of matching terms between a^+ and a^-. Hard negatives had $\mathrm{LexOverlap}(a^+, a^-)$ above a specific threshold yet were semantically distinct. This arrangement yielded a set of high-complexity pairs crucial for training a robust model.

We based our model on a Transformer encoder pretrained on a general medical corpus [67]. Let Θ represent the model parameters. For each training instance, we computed $\mathbf{h}(q)$, $\mathbf{h}(a^+)$, and $\mathbf{h}(a^-)$. The loss function combined a contrastive objective with a temperature term, given by [68]

$$\mathcal{L}(\Theta) = -\sum \log \frac{\exp(\mathrm{sim}(\mathbf{h}(q), \mathbf{h}(a^+))/\tau)}{\exp(\mathrm{sim}(\mathbf{h}(q), \mathbf{h}(a^+))/\tau) + \exp(\mathrm{sim}(\mathbf{h}(q), \mathbf{h}(a^-))/\tau)},$$

where τ is a temperature hyperparameter and $\mathrm{sim}(\cdot, \cdot)$ is the cosine similarity. The model was trained via stochastic gradient descent with a batch size that included an equal number of positive and negative examples. Validation runs were used to tune hyperparameters such as learning rate, batch size, and τ. [69]

Testing involved evaluating sentence similarity via the trained model on unseen question-answer pairs. We measured performance using precision at k (where k ranged from 1 to 10) and Mean Reciprocal Rank (MRR) [70]. The system was tasked with retrieving the correct answer a^+ for each question q from a set of M candidate answers, where M included both random and hard negatives [71]. Formally, if a_i^+ denotes the correct answer for question q_i, and the system ranks this answer at position r_i, the reciprocal rank is $1/r_i$. The MRR is then [72]

$$\mathrm{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r_i},$$

where N is the total number of test queries. A higher MRR signifies that correct answers generally appear near the top of the candidate list [73]. Precision at k tracks how many correct answers are found in the top k results.

Results demonstrated a marked improvement over standard embeddings such as a generic BERT model not fine-tuned for clinical text [74]. The contrastive model achieved an MRR of 0.82 for Diagnostic Queries and 0.86 for Treatment Queries, compared to 0.69 and 0.74, respectively, for the baseline. Precision at 1 was also notably higher, indicating that the first-ranked answer was correct a larger proportion of the time when using the contrastive embeddings [75]. This underscores the framework's ability to discriminate subtle domain differences, particularly in the presence of near-duplicate hard negatives.

Moreover, the results shed light on the efficacy of domain adaptation [76]. Models pretrained on medical text generally outperformed those pretrained solely on open-domain corpora, underscoring the domain mismatch phenomenon. An additional experiment evaluating the effect of adding unlabeled clinical notes to the pretraining data revealed further improvements [77]. The representation captured more fine-grained associations among medical terms, leading to a statistically significant increase in retrieval accuracy. Specifically, we observed an MRR increase of approximately 2.4

$$\Delta_{\text{MRR}} \propto \sqrt{|\text{Data}_{\text{in\_domain}}|}.$$

As the quantity of in-domain data grew, the marginal gains diminished, a phenomenon akin to the diminishing returns seen in general natural language processing tasks.

We also investigated the role of hyperparameter τ in controlling the softness of the contrastive distribution [78]. Too large a τ led to over-smoothing, while too small a τ caused unstable gradients. Optimal values for τ fell within the range $[0.05, 0.1]$, balancing the need for discriminating subtle differences in sentence meaning with stable training convergence [79]. Similarly, we explored different dimensionalities d for the final sentence embeddings. The best performance emerged at $d = 768$, which aligns with many Transformer-based architectures. [80]

Ablation studies attempted to clarify the importance of hard negatives. Removing them and relying solely on random negatives reduced MRR by around 3-4

Overall, these results validate the proposed approach, establishing that contrastive learning can drive meaningful gains in clinical sentence similarity estimation. By systematically moving correct question-answer pairs closer and pushing unrelated pairs apart, the framework achieves robust performance across diverse query types [81]. The next section delves into theoretical underpinnings that clarify why contrastive learning is particularly well-suited to capturing the manifold geometry of medical sentence embeddings. [82]

5. Theoretical Analysis of Sentence Encoding

The theoretical foundation for using contrastive learning in sentence embedding hinges on geometric principles in high-dimensional vector spaces. Medical texts, although idiosyncratic, can be viewed through the lens of manifold learning: semantically similar sentences are hypothesized to lie close to each other on a manifold embedded in $\mathbb{R}^d$. Contrastive learning imposes a structured penalty that reshapes this manifold to separate classes or clusters of sentences [83]. One can formalize this via topological and spectral arguments, illustrating how a properly tuned contrastive objective encourages favorable manifold properties, such as uniform coverage of valid semantic regions and distinct separations for dissimilar areas.

Consider an embedding function $\mathbf{h} : \mathcal{S} \to \mathbb{R}^d$. Let $\mathcal{M} \subset \mathbb{R}^d$ be the manifold that contains images of clinical sentences under $\mathbf{h}$. For a given semantic concept C, define $\mathcal{M}_C = \{\mathbf{h}(s) \mid s \in \mathcal{S}, \text{Sem}(s) = C\}$, where $\text{Sem}(s) = C$ indicates that sentence s encodes the concept C. Contrastive learning drives embeddings that keep each $\mathcal{M}_C$ as a compact submanifold and pushes it away from other submanifolds $\mathcal{M}_{C'}$ where $C' \neq C$. This separation can be cast in terms of the geodesic distance on $\mathcal{M}$. That is, if $d_{\mathcal{M}}$ denotes the intrinsic distance on $\mathcal{M}$, we aim for $d_{\mathcal{M}}(\mathbf{h}(s_1), \mathbf{h}(s_2))$ to be small for s_1, s_2 that share a concept and large otherwise.

Analyses of neural encoders often turn to Lipschitz continuity to bound how small changes in input textual tokens affect the resulting embeddings [84]. If $\mathbf{h}$ is L-Lipschitz, it satisfies:

$$\|\mathbf{h}(s_1) - \mathbf{h}(s_2)\| \leq L\,\mathrm{dist}_{\mathrm{token}}(s_1, s_2),$$

where $\mathrm{dist}_{\mathrm{token}}(\cdot,\cdot)$ measures some token-level edit or semantic distance. Contrastive learning, by enforcing margin constraints or distributional constraints, effectively places bounds on $\|\mathbf{h}(s_1) - \mathbf{h}(s_2)\|$ for various classes of sentence pairs. Through repeated gradient updates, it refines $\mathbf{h}$ to lower L in regions that correspond to frequent medical concepts, ensuring stable local neighborhoods.

A further perspective comes from spectral analysis of the Jacobian of $\mathbf{h}$. When $\mathbf{h}$ is parameterized by a deep neural network, consider the singular values of the Jacobian $J_{\mathbf{h}}(s)$. These values govern how stretching or compression in the input space is reflected in the output space. Minimizing a contrastive loss that punishes contradictory pairwise relationships can be interpreted as controlling the largest singular value, thereby limiting how drastically embeddings can move in response to small differences in input [85]. This ensures more uniform geometry. At the same time, the network retains enough expressive power to separate truly distinct concepts. [86]

In specialized domains like clinical medicine, sentences often cluster by pathology, treatment, or physiological measurements. One can define a set of equivalence relations among sentences: let $s_1 \sim_{\mathcal{E}} s_2$ denote that s_1 and s_2 belong to the same medical concept equivalence class $\mathcal{E}$. For example, "high blood pressure" may be considered semantically equivalent to "elevated BP." A well-trained contrastive model attempts to encode $\sim_{\mathcal{E}}$ as an equivalence relation in $\mathbb{R}^d$, meaning that $\mathbf{h}(s_1)$ and $\mathbf{h}(s_2)$ are neighbors if $s_1 \sim_{\mathcal{E}} s_2$. The required property is: [87]

$$s_1 \sim_{\mathcal{E}} s_2 \implies \|\mathbf{h}(s_1) - \mathbf{h}(s_2)\| \leq \epsilon$$

for some $\epsilon > 0$. Conversely, if s_1 and s_2 relate to different equivalences, the model tries to ensure [88]

$$s_1 \nsim_{\mathcal{E}} s_2 \implies \|\mathbf{h}(s_1) - \mathbf{h}(s_2)\| \geq \delta,$$

for some $\delta > 0$ with $\delta > \epsilon$. Contrastive losses serve as a computational proxy for these idealized constraints [89]. Although they cannot be perfectly enforced for all possible sentences, the repeated sampling of positive and negative pairs in training data approximates the coverage of crucial equivalence relations within medical language.

Another core element is the interpretability of these geometry-based constraints [90]. If the embedding space is well-structured, local neighborhoods can be consistently interpreted in terms of domain concepts. One might say that a local ball $B(\mathbf{h}(s_1), r)$ in $\mathbb{R}^d$ contains primarily sentences that correspond to the same or closely related medical concepts. This property underpins efficient retrieval: a nearest neighbor search for question embeddings can swiftly locate semantically appropriate answers [91]. Logic-based frameworks also dovetail with this perspective by linking the notion of semantic entailment to distance metrics. Under certain assumptions, an embedding-based method might represent entailment $\models$ as a containment relation between neighborhoods. [92]

Finally, from a complexity-theoretic standpoint, training such models is not trivial. The gradient-based updates must navigate a high-dimensional parameter space [93]. Although a rigorous proof of convergence in non-convex settings remains elusive, empirical evidence and partial theoretical arguments support the conclusion that repeated exposure to diverse positive and negative pairs systematically shapes $\mathbf{h}$. One may argue that for large-scale training data drawn from the true distribution of medical sentences, the model eventually learns a stable manifold separation that generalizes effectively to new queries. The success of the approach in practice, as described in prior sections, provides evidence of these manifold assumptions holding to a satisfactory extent. [94]

These theoretical insights offer a deeper understanding of why contrastive learning is particularly adept at capturing the complexities of medical sentence embeddings [95]. The interplay between geometric constraints, Lipschitz regularity, and domain-specific equivalence relations culminates in an embedding space supportive of high-performing question answering. Given the synergy between theory and practice demonstrated, we now turn our attention to summarizing the key contributions and potential future directions. [96]

6. Conclusion

In this paper, we presented a deep exploration into the role of contrastive learning for estimating clinical sentence similarity in medical question answering systems. By focusing on domain-specific subtleties such as specialized abbreviations, negation handling, and context-sensitive interpretations of medical terminology, we constructed an embedding space conducive to accurate query-to-answer matching [97]. The central premise was that contrastive objectives, which inherently reward proximity for semantically aligned sentence pairs and enforce distance for unrelated pairs, can yield robust embeddings despite the inherent noise and complexity of clinical text.

We began with a discussion of the essential challenges in medical language understanding, emphasizing the unique properties of clinical narratives that set them apart from general-domain corpora [98]. These include the prevalence of partial or inconsistent data, the sensitivity of numeric values and measurements, and the high stakes associated with misinterpretations in a healthcare environment. We then explored the structural framework of contrastive learning, detailing how margin-based or log-based objectives guide the model toward nuanced discrimination of sentence semantics [99]. Empirical evaluations on a variety of clinical question-answer datasets affirmed the potential of this approach, showing substantial gains in metrics like Mean Reciprocal Rank and precision at k over baseline methods. Notably, experiments confirmed that domain-adaptive pretraining and the strategic use of hard negative examples were instrumental in driving performance improvements. [100, 101]

We additionally highlighted theoretical considerations, framing medical sentence embeddings in terms of manifold geometry, Lipschitz continuity, and logical equivalence relations. These formulations demystified some of the mechanics by which contrastive learning shapes the embedding space, underscoring the importance of local neighborhood structures in capturing subtle distinctions among clinical concepts [102]. While the formal proofs of convergence in non-convex neural network training remain out of reach, empirical evidence and partial analyses provide strong support that repeated exposure to semantically aligned and unaligned sentence pairs can effectively reorganize the manifold of representations to reflect domain-specific concepts.

Looking forward, there are multiple avenues for advancing the techniques discussed here [103]. One direction involves refining the construction of hard negative examples by incorporating knowledge graphs or specialized ontologies that systematically identify near-synonyms or closely related conditions. Another direction is to explore multi-lingual medical corpora, expanding the application of contrastive learning to globally diverse healthcare contexts [104]. Because the medical domain often spans multiple languages and regions, bridging these linguistic gaps could significantly broaden the utility of question answering systems. A third direction involves the integration of structured data sources like lab measurements and imaging results, fusing textual embeddings with multimodal inputs to formulate even richer representations [105]. The potential benefits range from improved specificity in question answering to the discovery of latent correlations between text, lab findings, and clinical outcomes.

Lastly, ethical and privacy considerations remain an overarching concern in clinical NLP [106]. The processes for collecting, annotating, and sharing patient data require rigorous adherence to confidentiality standards. Any practical system must incorporate data de-identification and anonymization protocols, potentially limiting the data available for model pretraining [107]. Nevertheless, the modular nature of contrastive learning frameworks can accommodate these constraints by tailoring objectives to partially labeled or masked datasets.

In summary, the methodological, empirical, and theoretical insights gleaned from this work collectively underscore the viability of contrastive learning for clinical sentence similarity estimation [108]. Through the lens of both experimental outcomes and analytic reasoning, we have demonstrated how mapping semantically related sentences to close-by representations yields tangible performance benefits for medical question answering. We anticipate that continued research in contrastive methods, especially when integrated with emerging multi-modal and knowledge graph capabilities, will further enhance the fidelity and practicality of intelligent systems in healthcare. [109]

References

[1] S.-F. Sung, K. Chen, D. P. Wu, L.-C. Hung, Y. H. Su, and Y. H. Hu, “Applying natural language processing techniques to develop a task-specific emr interface for timely stroke thrombolysis: A feasibility study.,” *International journal of medical informatics*, vol. 112, pp. 149–157, 2 2018.

[2] S. S. Doerstling, D. Akrobetu, M. M. Engelhard, F. Chen, and P. A. Ubel, “A disease identification algorithm for medical crowdfunding campaigns: Validation study.,” *Journal of medical Internet research*, vol. 24, pp. e32867–e32867, 6 2022.

[3] A. A. T. Monfared, Y. Stern, S. Doogan, M. Irizarry, and Q. Zhang, “Understanding barriers along the patient journey in alzheimer’s disease using social media data.,” *Neurology and therapy*, vol. 12, pp. 899–918, 4 2023.

[4] K. Joseph, H.-Y. W. Chen, S. Ionescu, Y. Du, P. Sankhe, A. Hannak, and A. Rudra, “A qualitative, network-centric method for modeling socio-technical systems, with applications to evaluating interventions on social media platforms to increase social equality,” *Applied Network Science*, vol. 7, 7 2022.

[5] K. D. Forbus, C. Riesbeck, L. Birnbaum, K. Livingston, A. Sharma, and L. Ureel, “Integrating natural language, knowledge representation and reasoning, and analogical processing to learn by reading,” in *Proceedings of the National Conference on Artificial Intelligence*, vol. 22, p. 1542, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[6] J. A. M. Sidey-Gibbons and C. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC medical research methodology*, vol. 19, pp. 64–64, 3 2019.

[7] M. Yuan and A. Vlachos, “Zero-shot fact-checking with semantic triples and knowledge graphs,” in *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pp. 105–115, 2024.

[8] Y. Wang, H. Xu, and Özlem Uzuner, “Editorial: The second international workshop on health natural language processing (healthnlp 2019).,” *BMC medical informatics and decision making*, vol. 19, pp. 1–3, 12 2019.

[9] S. Storey, X. Luo, S. Ofner, S. M. Perkins, and D. V. Ah, “Hyperglycemia, symptoms, and symptom clusters in colorectal cancer survivors with type 2 diabetes.,” *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*, vol. 30, pp. 10149–10157, 11 2022.

[10] A. H. Zhao, D. I. Glazer, M. M. Hammer, K. S. Burk, P. J. DiPiro, and R. Khorasani, “Comparing thoracic and abdominal subspecialists’ follow-up recommendations for abdominal findings identified on chest ct.,” *Abdominal radiology (New York)*, vol. 48, pp. 1468–1478, 2 2023.

[11] I. D. Dinov, “Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data,” *GigaScience*, vol. 5, pp. 12–12, 2 2016.

[12] S. Kang, L. Patil, A. Rangarajan, A. Moitra, T. Jia, D. Robinson, F. Ameri, and D. Dutta, “Extraction of formal manufacturing rules from unstructured english text,” *Computer-Aided Design*, vol. 134, pp. 102990–, 2021.

[13] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, “Clinical natural language processing in languages other than english: opportunities and challenges,” *Journal of biomedical semantics*, vol. 9, pp. 12–12, 3 2018.

[14] E. Mayes, J. A. Gehlbach, P. M. Jeziorczak, and A. R. Wooldridge, “Machine learning to operationalize team cognition: A case study of patient handoffs,” *Human Factors in Healthcare*, vol. 3, pp. 100036–100036, 2023.

[15] M. I. Miller, L. C. Shih, and V. B. Kolachalama, “Machine learning in clinical trials: A primer with applications to neurology.,” *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics*, vol. 20, pp. 1066–1080, 5 2023.

[16] L. Neely, A. Carnett, J. Quarles, H. MacNaul, S.-W. Park, S. Oyama, G. Chen, K. Desai, and P. Najafirad, “The case for integrated advanced technology in applied behavior analysis,” *Advances in Neurodevelopmental Disorders*, vol. 7, pp. 415–425, 12 2022.

[17] M. Abouelyazid and C. Xiang, “Machine learning-assisted approach for fetal health status prediction using cardiotocogram data,” *International Journal of Applied Health Care Analytics*, vol. 6, no. 4, pp. 1–22, 2021.

[18] M. Salas, J. Petracek, P. Yalamanchili, O. Aimer, D. Kasthuril, S. Dhingra, T. Junaid, and T. Bostic, “The use of artificial intelligence in pharmacovigilance: A systematic review of the literature.,” *Pharmaceutical medicine*, vol. 36, pp. 295–306, 7 2022.

[19] W. M. Perry, R. Hossain, and R. A. Taylor, “Assessment of the feasibility of automated, real-time clinical decision support in the emergency department using electronic health record data,” *BMC emergency medicine*, vol. 18, pp. 19–19, 7 2018.

[20] E. Bett, T. C. Frommeyer, T. Reddy, and J. Johnson, “Assessment of patient perceptions of technology and the use of machine-based learning in a clinical encounter,” *SSRN Electronic Journal*, 1 2022.

[21] Y. Gao, D. Dligach, T. Miller, M. M. Churpek, O. Uzuner, and M. Afshar, “Progress note understanding - assessment and plan reasoning: Overview of the 2022 n2c2 track 3 shared task.,” *Journal of biomedical informatics*, vol. 142, pp. 104346–104346, 4 2023.

[22] V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox, “Assessing the impact of a health intervention via user-generated internet content,” *Data Mining and Knowledge Discovery*, vol. 29, pp. 1434–1457, 7 2015.

[23] J. R. Machireddy, “Harnessing ai and data analytics for smarter healthcare solutions,” *International Journal of Science and Research Archive*, vol. 08, no. 02, pp. 785–798, 2023.

[24] R. Gerard, V. Makeeva, B. Vey, T. S. Cook, P. Nagy, R. W. Filice, K. C. Wang, P. Balthazar, P. Harri, and N. M. Safdar, “Imaging informatics fellowship curriculum: Building consensus on the most critical topics and the future of the informatics fellowship.,” *Journal of digital imaging*, vol. 36, pp. 1–10, 10 2022.

[25] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya, A. Gelbukh, and R. Mihalcea, “Recognizing emotion cause in conversations,” *Cognitive Computation*, vol. 13, pp. 1317–1332, 9 2021.

[26] A. Vaid, E. Argulian, S. Lerakis, B. K. Beaulieu-Jones, C. Krittanawong, E. Klang, J. Lampert, V. Y. Reddy, J. Narula, G. N. Nadkarni, and B. S. Glicksberg, “Multi-center retrospective cohort study applying deep learning to electrocardiograms to identify left heart valvular dysfunction.,” *Communications medicine*, vol. 3, pp. 24–, 2 2023.

[27] J. D. Gaizo, J. S. Obeid, K. R. Catchpole, and A. V. Alekseyenko, “Red flag/blue flag visualization of a common cnn for text classification.,” *JAMIA open*, vol. 6, pp. ooac112–, 1 2023.

[28] P. Parmar, J. Ryu, S. Pandya, J. Sedoc, and S. Agarwal, “Health-focused conversational agents in person-centered care: a review of apps.,” *NPJ digital medicine*, vol. 5, pp. 21–, 2 2022.

[29] J. H. Garvin, P. L. Elkin, S. Shen, S. H. Brown, B. Trusko, E. Wang, L. Hoke, Y. Quiaoit, J. LaJoie, M. G. Weiner, P. Graham, and T. Speroff, “Automated quality measurement in department of the veterans affairs discharge instructions for patients with congestive heart failure.,” *Journal for healthcare quality : official publication of the National Association for Healthcare Quality*, vol. 35, no. 4, pp. 16–24, 2013.

[30] A. K. Saxena, “Evaluating the regulatory and policy recommendations for promoting information diversity in the digital age,” *International Journal of Responsible Artificial Intelligence*, vol. 11, no. 8, pp. 33–42, 2021.

[31] M. McLenon, S. Okuhn, E. M. Lancaster, M. M. Hull, J. L. Adams, E. A. McGlynn, A. L. Avins, and R. W. Chang, “Validation of natural language processing to determine the presence and size of abdominal aortic aneurysms in a large integrated health system.,” *Journal of vascular surgery*, vol. 74, pp. 459–466.e3, 2 2021.

[32] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, “Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond,” *Knowledge and Information Systems*, vol. 64, pp. 3197–3234, 9 2022.

[33] Z. Peng, P. Rathod, N. Niu, T. Bhowmik, H. Liu, L. Shi, and Z. Jin, “Testing software’s changing features with environment-driven abstraction identification.,” *Requirements engineering*, vol. 27, pp. 405–427, 9 2022.

[34] C. Micale, S. Golder, K. O’Connor, D. Weissenbacher, R. Gross, S. Hennessy, and G. Gonzalez-Hernandez, “Patient-reported reasons for antihypertensive medication change: A quantitative study using social media.,” *Drug safety*, vol. 47, pp. 81–91, 11 2023.

[35] C. Yuan and S. S. Agaian, “A comprehensive review of binary neural network,” *Artificial Intelligence Review*, vol. 56, pp. 12949–13013, 3 2023.

[36] A. Wismüller, A. M. DSouza, A. Z. Abidin, M. A. Vosoughi, C. Gange, I. O. Cortopassi, G. Bozovic, A. A. Bankier, K. Batra, Y. Chodakiewitz, Y. Xi, C. T. Whitlow, J. Ponnatapura, G. J. Wendt, E. P. Weinberg, L. Stockmaster, D. A. Shrier, M. C. Shin, R. Modi, H. S. Lo, S. Kligerman, A. Hamid, L. D. Hahn, G. M. Garcia, J. H. Chung, T. Altes, S. Abbara, and A. S. Bader, "Early-stage covid-19 pandemic observations on pulmonary embolism using nationwide multi-institutional data harvesting.," *NPJ digital medicine*, vol. 5, pp. 120–, 8 2022.

[37] L. S. Gold, R. F. Cody, W. K. Tan, Z. A. Marcum, E. N. Meier, K. J. Sherman, K. T. James, B. Griffith, A. L. Avins, D. F. Kallmes, P. Suri, J. L. Friedly, P. J. Heagerty, R. A. Deyo, P. H. Luetmer, S. D. Rundell, D. R. Haynor, and J. G. Jarvik, "Osteoporosis identification among previously undiagnosed individuals with vertebral fractures.," *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 33, pp. 1925–1935, 6 2022.

[38] B. Kompa, J. B. Hakim, A. Palepu, K. G. Kompa, M. Smith, P. A. Bain, S. Woloszynek, J. L. Painter, A. Bate, and A. L. Beam, "Artificial intelligence based on machine learning in pharmacovigilance: A scoping review.," *Drug safety*, vol. 45, pp. 477–491, 5 2022.

[39] V. Roche, J.-P. Robert, and H. Salam, "Ai-based approach for safety signals detection from social networks: Application to the levothyrox scandal in 2017 on doctissimo forum," *SSRN Electronic Journal*, 1 2021.

[40] A. Sharma and K. M. Goolsbey, "Simulation-based approach to efficient commonsense reasoning in very large knowledge bases," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 1360–1367, 2019.

[41] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Approximate query processing for big data in heterogeneous databases," in *2020 IEEE international conference on big data (big data)*, pp. 5765–5767, IEEE, 2020.

[42] D. Kim, J. Chung, J. Choi, M. D. Succi, J. Conklin, M. G. F. Longo, J. B. Ackman, B. P. Little, M. Petranovic, M. K. Kalra, M. H. Lev, and S. Do, "Accurate auto-labeling of chest x-ray images based on quantitative similarity to an explainable ai model.," *Nature communications*, vol. 13, pp. 1867–, 4 2022.

[43] R. Avula, "Architectural frameworks for big data analytics in patient-centric healthcare systems: Opportunities, challenges, and limitations," *Emerging Trends in Machine Intelligence and Big Data*, vol. 10, no. 3, pp. 13–27, 2018.

[44] A. Maghsoudi, Y. H. Sada, H. Zhu, S. Yarlagadda, D. Guffey, S. Nowakowski, A. Li, and J. Razjouyan, "Application of natural language processing to assess the performance status documentation quality metric in patients with non–small-cell lung cancer.," *Journal of Clinical Oncology*, vol. 41, pp. e13582–e13582, 6 2023.

[45] Z. Li, K. Hwang, K. Li, J. Wu, and T. Ji, "Graph-generative neural network for eeg-based epileptic seizure detection via discovery of dynamic brain functional connectivity.," *Scientific reports*, vol. 12, pp. 18998–, 11 2022.

[46] W. Zhong, P. Y. Yao, S. H. Boppana, F. V. Pacheco, B. S. Alexander, S. Simpson, and R. A. Gabriel, "Improving case duration accuracy of orthopedic surgery using bidirectional encoder representations from transformers (bert) on radiology reports.," *Journal of clinical monitoring and computing*, vol. 38, pp. 221–228, 9 2023.

[47] H. Piwowar and W. Chapman, "Identifying data sharing in biomedical literature," *Nature Precedings*, 3 2008.

[48] N. Rashid, G. Levy, Y.-L. Wu, C. Zheng, R. Koblick, and T. C. Cheetham, "Patient and clinical characteristics associated with gout flares in an integrated healthcare system.," *Rheumatology international*, vol. 35, pp. 1799–1807, 5 2015.

[49] S. Khurshid, C. Reeder, L. X. Harrington, P. Singh, G. Sarma, S. F. Friedman, P. D. Achille, N. Diamant, J. W. Cunningham, A. C. Turner, E. S. Lau, J. S. Haimovich, M. A. Al-Alusi, X. Wang, M. D. R. Klarqvist, J. M. Ashburner, C. Diedrich, M. Ghadessi, J. Mielke, H. M. Eilken, A. McElhinney, A. Derix, S. J. Atlas, P. T. Ellinor, A. A. Philippakis, C. D. Anderson, J. E. Ho, P. Batra, and S. A. Lubitz, "Cohort design and natural language processing to reduce bias in electronic health records research.," *NPJ digital medicine*, vol. 5, pp. 47–, 4 2022.

[50] A. L. Beam, U. Kartoun, J. K. Pai, A. K. Chatterjee, T. P. Fitzgerald, S. Y. Shaw, and I. S. Kohane, "Predictive modeling of physician-patient dynamics that influence sleep medication prescriptions and clinical decision-making.," *Scientific reports*, vol. 7, pp. 42282–42282, 2 2017.

[51] J. A. Rodger and J. Piper, "Assessing american presidential candidates using principles of ontological engineering, word sense disambiguation, data envelope analysis and qualitative comparative analysis," *International Journal of Speech Technology*, vol. 26, pp. 743–764, 10 2023.

[52] J. L. Pease, D. Thompson, J. Wright-Berryman, and M. Campbell, "User feedback on the use of a natural language processing application to screen for suicide risk in the emergency department.," *The journal of behavioral health services & research*, vol. 50, pp. 548–554, 2 2023.

[53] Q. Lu, Z. Cui, Y. Chen, and X. Chen, “Extracting optimal actionable plans from additive tree models,” *Frontiers of Computer Science*, vol. 11, pp. 160–173, 4 2017.

[54] A. Minhajuddin, M. K. Jha, C. R. C. Fatt, T. L. Mayes, J. D. Berry, M. H. Trivedi, J. Dennison, C. Volmar, J. A. Timmons, C. Wahlestedt, A. A. Keiser, T. Dong, E. A. Kramár, C. Butler, D. P. Matheos, L. Tong, N. Berchtold, S. Chen, M. Samad, J. Beardwood, S. Shanur, A. M. Rodriguez, P. Baldi, C. W. Cotman, M. A. Wood, M. Ananth, D. A. Talmage, V. K. Martinez, B. McAlpin, R. Mahalingam, I. Mahant, A. Kavelaars, and C. J. Heijnen, “Acnp 60th annual meeting: Poster abstracts p1 - p275.,” *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, vol. 46, pp. 72–217, 12 2021.

[55] V. Bali, J. Weaver, V. Turzhitsky, J. Schelfhout, M. L. Paudel, E. Hulbert, J. Peterson-Brandt, A.-M. G. Currie, and D. Bakka, “Development of a natural language processing algorithm to detect chronic cough in electronic health records.,” *BMC pulmonary medicine*, vol. 22, pp. 256–, 6 2022.

[56] L. Anzaldi, A. Davison, C. M. Boyd, B. Leff, and H. Kharrazi, “Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study.,” *BMC geriatrics*, vol. 17, pp. 248–248, 10 2017.

[57] M. Vithayathil, S. Smith, S. Goryachev, J. Nayor, and M. Song, “Development of a large colonoscopy-based longitudinal cohort for integrated research of colorectal cancer: Partners colonoscopy cohort.,” *Digestive diseases and sciences*, vol. 67, pp. 1–8, 2 2021.

[58] L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, “Machine learning for medical ultrasound: status, methods, and future opportunities.,” *Abdominal radiology (New York)*, vol. 43, pp. 786–799, 2 2018.

[59] R. H. Perlis, D. V. Iosifescu, V. M. Castro, S. N. Murphy, V. S. Gainer, J. Minnier, T. Cai, S. Goryachev, Q. Zeng, P. J. Gallagher, M. Fava, J. B. Weilburg, S. Churchill, I. S. Kohane, and J. W. Smoller, “Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model,” *Psychological medicine*, vol. 42, pp. 41–50, 6 2011.

[60] C. Zheng, B. C. Sun, Y.-L. Wu, M. Ferencik, M.-S. Lee, R. F. Redberg, A. A. Kawatkar, V. V. Musigdilok, and A. L. Sharp, “Automated interpretation of stress echocardiography reports using natural language processing.,” *European heart journal. Digital health*, vol. 3, pp. 626–637, 9 2022.

[61] J. P. Ridgway, A. Lee, S. Devlin, J. Kerman, and A. Mayampurath, “Machine learning and clinical informatics for improving hiv care continuum outcomes.,” *Current HIV/AIDS reports*, vol. 18, pp. 229–236, 3 2021.

[62] L. Cui, X. Xie, and Z.-J. M. Shen, “Prediction task guided representation learning of medical codes in ehr.,” *Journal of biomedical informatics*, vol. 84, pp. 1–10, 6 2018.

[63] M. Giuffrè and D. L. Shung, “Harnessing the power of synthetic data in healthcare: innovation, application, and privacy.,” *NPJ digital medicine*, vol. 6, pp. 186–, 10 2023.

[64] D. Vetter, J. Amann, F. Bruneault, M. Coffee, B. Düdder, A. Gallucci, T. K. Gilbert, T. Hagendorff, I. van Halem, E. Hickman, E. Hildt, S. Holm, G. Kararigas, P. Kringen, V. I. Madai, E. W. Mathez, J. J. Tithi, M. Westerlund, R. Wurth, and R. V. Zicari, “Lessons learned from assessing trustworthy ai in practice,” *Digital Society*, vol. 2, 9 2023.

[65] A. Lebal, A. Moussaoui, and A. Rezgui, “Epilepsy-net: attention-based 1d-inception network model for epilepsy detection using one-channel and multi-channel eeg signals,” *Multimedia Tools and Applications*, vol. 82, pp. 17391–17413, 10 2022.

[66] A. E. Levy, N. R. Shah, M. E. Matheny, R. M. Reeves, G. T. Gobbel, and S. M. Bradley, “Determining post-test risk in a national sample of stress nuclear myocardial perfusion imaging reports: Implications for natural language processing tools.,” *Journal of nuclear cardiology : official publication of the American Society of Nuclear Cardiology*, vol. 26, pp. 1878–1885, 4 2018.

[67] J. Wong, D. Prieto-Alhambra, P. R. Rijnbeek, R. J. Desai, J. M. Reps, and S. Toh, “Applying machine learning in distributed data networks for pharmacoepidemiologic and pharmacovigilance studies: Opportunities, challenges, and considerations.,” *Drug safety*, vol. 45, pp. 493–510, 5 2022.

[68] M. A. Al-Garadi, Y.-C. Yang, and A. Sarker, “The role of natural language processing during the covid-19 pandemic: Health applications, opportunities, and challenges.,” *Healthcare (Basel, Switzerland)*, vol. 10, pp. 2270–2270, 11 2022.

[69] L. J. J, K. Singh, and B. Chakravarthi, “Digital forensic framework for smart contract vulnerabilities using ensemble models,” *Multimedia Tools and Applications*, vol. 83, pp. 51469–51512, 11 2023.

[70] D. Torre, F. Mesadieu, and A. Chennamaneni, “Deep learning techniques to detect cybersecurity attacks: a systematic mapping study,” *Empirical Software Engineering*, vol. 28, 5 2023.

[71] B. T. Bucher, J. Shi, J. P. Ferraro, D. E. Skarda, M. H. Samore, J. F. Hurdle, A. V. Gundlapalli, W. W. Chapman, and S. R. Finlayson, “Portable automated surveillance of surgical site infections using natural language processing: Development and validation.,” *Annals of surgery*, vol. 272, pp. 629–636, 7 2020.

[72] R. Avula, “Addressing barriers in data collection, transmission, and security to optimize data availability in healthcare systems for improved clinical decision-making and analytics,” *Applied Research in Artificial Intelligence and Cloud Computing*, vol. 4, no. 1, pp. 78–93, 2021.

[73] A. Vaid, J. Jiang, A. Sawant, S. Lerakis, E. Argulian, Y. Ahuja, J. Lampert, A. Charney, H. Greenspan, J. Narula, B. Glicksberg, and G. N. Nadkarni, “A foundational vision transformer improves diagnostic performance for electrocardiograms.,” *NPJ digital medicine*, vol. 6, pp. 108–, 6 2023.

[74] M. C. Beach, S. Saha, J. Park, J. L. Taylor, P. Drew, E. Plank, L. A. Cooper, and B. W. Chee, “Testimonial injustice: Linguistic bias in the medical records of black patients and women,” *Journal of general internal medicine*, vol. 36, pp. 1708–1714, 3 2021.

[75] J. Jang, A. A. Colletti, C. Ricklefs, H. J. Snyder, K. Kardonsky, E. W. Duggan, G. E. Umpierrez, and V. N. O’Reilly-Shah, “Implementation of app-based diabetes medication management: Outpatient and perioperative clinical decision support.,” *Current diabetes reports*, vol. 21, pp. 50–, 12 2021.

[76] B. I. Reiner, “Using analysis of speech and linguistics to characterize uncertainty in radiology reporting,” *Journal of digital imaging*, vol. 25, pp. 703–707, 10 2012.

[77] V. Barcelona, D. Scharp, H. Moen, A. Davoudi, B. R. Idnay, K. Cato, and M. Topaz, “Using natural language processing to identify stigmatizing language in labor and birth clinical notes.,” *Maternal and child health journal*, vol. 28, pp. 578–586, 12 2023.

[78] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “Federated query processing for big data in data science,” in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 6145–6147, IEEE, 2019.

[79] E. Kellar, S. M. Bornstein, A. Caban, M. Crouthamel, C. Celingant, P. A. McIntire, C. Johnson, P. Mehta, V. Sikirica, and B. Wilson, “Optimizing the use of electronic data sources in clinical trials: The technology landscape.,” *Therapeutic innovation & regulatory science*, vol. 51, pp. 551–567, 7 2017.

[80] D. Liu, D. Dligach, and T. A. Miller, “Bionlp@acl - two-stage federated phenotyping and patient representation learning.,” *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, pp. 283–291, 2019.

[81] J. M. Boggs, A. Beck, D. P. Ritzwoller, C. Battaglia, H. D. Anderson, and R. C. Lindrooth, “A quasi-experimental analysis of lethal means assessment and risk for subsequent suicide attempts and deaths,” *Journal of general internal medicine*, vol. 35, pp. 1709–1714, 2 2020.

[82] Y. R. Patel, J. M. Robbins, K. E. Kurgansky, T. F. Imran, A. R. Orkaby, R. R. McLean, Y.-L. Ho, K. Cho, J. M. Gaziano, L. Djoussé, D. R. Gagnon, and J. Joseph, “Development and validation of a heart failure with preserved ejection fraction cohort using electronic medical records.,” *BMC cardiovascular disorders*, vol. 18, pp. 128–128, 6 2018.

[83] Z. N. Kiss, K. Bogos, L. Tamási, G. Ostoros, V. Müller, N. Bittner, V. Sárosi, A. Vastag, K. Knollmajer, M. Várnai, K. Kovács, A. Berta, I. Köveskuti, E. Karamousouli, G. Rokszin, Z. Abonyi-Tóth, Z. Barcza, I. Kenessey, A. Weber, P. Nagy, P. Freyler-Fadgyas, M. Szócska, P. Szegner, L. Hilbert, G. B. Géczy, G. Surján, J. Moldvay, Z. Vokó, G. Gálffy, and Z. Polányi, “Underlying reasons for post-mortem diagnosed lung cancer cases - a robust retrospective comparative study from hungary (hulc study).,” *Frontiers in oncology*, vol. 12, pp. 1032366–, 11 2022.

[84] N. Tavabi, D. Stück, A. Signorini, C. Karjadi, T. A. Hanai, M. Sandoval, C. Lemke, J. Glass, S. Hardy, M. Lavallee, B. Wasserman, T. F. A. Ang, C. M. Nowak, R. Kainkaryam, L. Foschini, and R. Au, “Cognitive digital biomarkers from automated transcription of spoken language.,” *The journal of prevention of Alzheimer’s disease*, vol. 9, no. 4, pp. 791–800, 2022.

[85] K. Ostherr, “Artificial intelligence and medical humanities.,” *The Journal of medical humanities*, vol. 43, pp. 1–22, 7 2020.

[86] A. Sharma and K. D. Forbus, “Modeling the evolution of knowledge and reasoning in learning systems,” in *2010 AAAI Fall Symposium Series*, 2010.

[87] S. Bhatt, P. C. Johnson, N. H. Markovitz, T. Gray, R. D. Nipp, N. Ufere, J. Rice, M. J. Reynolds, M. W. Lavoie, M. A. Clay, C. Lindvall, and A. El-Jawahri, "The use of natural language processing to assess social support in patients with advanced cancer.," *The oncologist*, vol. 28, pp. 165–171, 11 2022.

[88] J. Han, K. Chen, L. Fang, S. Zhang, F. Wang, H. Ma, L. Zhao, and S. Liu, "Improving the efficacy of the data entry process for clinical research with a natural language processing-driven medical information extraction system: Quantitative field research.," *JMIR medical informatics*, vol. 7, pp. e13331–, 7 2019.

[89] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural language processing for mental health interventions: a systematic review and research framework.," *Translational psychiatry*, vol. 13, pp. 309–, 10 2023.

[90] S. T. Garrity, M. Pistilli, M. S. Vaphiades, N. Q. Richards, P. S. Subramanian, P. R. Rosa, B. L. Lam, B. Osborne, G. T. Liu, K. E. Duncan, R. K. Shin, N. J. Volpe, K. S. Shindler, M. S. Lee, M. L. Moster, E. H. Tracey, S. E. Cuprill-Nilson, and M. A. Tamhankar, "Ophthalmic presentation of giant cell arteritis in african-americans," *Eye (London, England)*, vol. 31, pp. 113–118, 9 2016.

[91] I.-E. Nogues, J. Wen, Y. Lin, M. Liu, S. K. Tedeschi, A. Geva, T. Cai, and C. Hong, "Weakly semi-supervised phenotyping using electronic health records.," *Journal of biomedical informatics*, vol. 134, pp. 104175–104175, 9 2022.

[92] K. Smith, S. Golder, A. Sarker, Y. K. Loke, K. O'Connor, and G. Gonzalez-Hernandez, "Methods to compare adverse events in twitter to faers, drug information databases, and systematic reviews: Proof of concept with adalimumab," *Drug safety*, vol. 41, pp. 1397–1410, 8 2018.

[93] Y. Dang, F. Li, X. Hu, V. K. Keloth, M. Zhang, S. Fu, M. F. Amith, J. W. Fan, J. Du, E. Yu, H. Liu, X. Jiang, H. Xu, and C. Tao, "Systematic design and data-driven evaluation of social determinants of health ontology (sdoho).," *Journal of the American Medical Informatics Association : JAMIA*, vol. 30, pp. 1465–1473, 6 2023.

[94] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC medical informatics and decision making*, vol. 17, pp. 1–13, 12 2017.

[95] N. Wakutsu, E. Hirose, N. Yonemoto, and S. Demiya, "Assessing definitions and incentives adopted for innovation for pharmaceutical products in five high-income countries: A systematic literature review.," *Pharmaceutical medicine*, vol. 37, pp. 53–70, 1 2023.

[96] A. Sharma and K. D. Forbus, "Graph-based reasoning and reinforcement learning for improving q/a performance in large knowledge-based systems," in *2010 AAAI Fall Symposium Series*, 2010.

[97] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review," *Journal of the American Medical Informatics Association : JAMIA*, vol. 26, pp. 364–379, 2 2019.

[98] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Automatic visual recommendation for data science and analytics," in *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*, pp. 125–132, Springer, 2020.

[99] Y.-H. Sheu, C. Magdamo, M. Miller, S. Das, D. Blacker, and J. W. Smoller, "Ai-assisted prediction of differential response to antidepressant classes using electronic health records.," *NPJ digital medicine*, vol. 6, pp. 73–, 4 2023.

[100] A. Liede, W. Sebby, A. K. R. Miriyala, R. Potluri, D. Mazumder, A. Ghosh, E. Papademetriou, R. Kilpatrick, and J. E. Tyczynski, "Risk of seizures in a population of women with brca-positive metastatic breast cancer from an electronic health record database in the united states.," *BMC cancer*, vol. 23, pp. 78–, 1 2023.

[101] J. R. Machireddy, "Automation in healthcare claims processing: Enhancing efficiency and accuracy," *International Journal of Science and Research Archive*, vol. 09, no. 01, pp. 825–834, 2023.

[102] R. Mylvaganam, R. Lawrence, I. Goldberg, F. Rahaghi, S. Chiu, S. C. Malaisrie, D. Schimmel, R. Avery, K. Martin, and M. J. Cuttica, "Differences in referral to a chronic thromboembolic pulmonary hypertension center following acute pulmonary embolism: a locoregional experience.," *Journal of thrombosis and thrombolysis*, vol. 55, pp. 691–699, 2 2023.

[103] S. H. Lee, "Natural language generation for electronic health records.," *NPJ digital medicine*, vol. 1, pp. 63–63, 11 2018.

[104] A. Telenti and X. Jiang, "Treating medical data as a durable asset.," *Nature genetics*, vol. 52, pp. 1005–1010, 9 2020.

[105] R. Avula, “Applications of bayesian statistics in healthcare for improving predictive modeling, decision-making, and adaptive personalized medicine,” *International Journal of Applied Health Care Analytics*, vol. 7, no. 11, pp. 29–43, 2022.

[106] F. M. Asch, R. P. Sharma, R. J. Cubeddu, P. Généreux, M. Dobbles, K. Verhoef, M. Kwon, E. Rodriguez, J. D. Thomas, and L. D. Gillam, “Gaps in contemporary echocardiographic reporting quality for mechanisms of mitral regurgitation: A call to action.,” *Journal of the American Society of Echocardiography : official publication of the American Society of Echocardiography*, vol. 37, pp. 108–110, 9 2023.

[107] C. Xiang and M. Abouelyazid, “The impact of generational cohorts and visit environment on telemedicine satisfaction: A novel investigation,” 2020.

[108] B. Cade, S. Hassan, H. Dashti, M. Kiernan, M. Pavlova, S. Redline, and E. Karlson, “0700 prospective and cross-sectional associations between sleep apnea and disease in a phenome-wide analysis of a clinical biobank,” *Sleep*, vol. 45, pp. A307–A307, 5 2022.

[109] A. T. Bako, H. L. Taylor, K. K. Wiley, J. Zheng, H. Walter-McCabe, S. N. Kasthurirathne, and J. R. Vest, “Using natural language processing to classify social work interventions.,” *The American journal of managed care*, vol. 27, pp. e24–e31, 1 2021.