ORIGINAL RESEARCH **Open access**

A Multimodal Foundation Model for Zero-Shot Rare Disease Diagnosis from Electronic Health Records

Ahmed Youssef[1*], Khaled Hassan[1], Mahmoud Elamin[2]

Abstract

Rare diseases collectively affect over 300 million people globally, yet individual conditions are often missed due to low clinician familiarity and non-specific presenting symptoms that mimic common disorders. Supervised machine learning requires large numbers of labeled examples for training, but rare diseases have too few diagnosed cases to develop condition-specific predictive models using traditional approaches. We propose a multimodal foundation model pretrained on 10 million de-identified electronic health records (EHRs) combining clinical notes and laboratory values for zero-shot rare disease diagnosis without requiring labeled training examples. The framework comprises four components: a clinical note encoder based on a large language model, a laboratory value encoder using a time-series transformer, a multimodal fusion module with cross-attention, and a zero-shot classifier that compares patient embeddings to disease descriptions. Pretraining on large-scale EHR data enables the model to learn general medical knowledge and disease patterns, allowing diagnosis of rare conditions by recognizing manifestations even when no labeled examples of that specific disease were used for training.

Keywords Foundation models, Electronic health records, Zero-shot learning, Clinical NLP, Rare diseases, Multimodal learning

*Correspondence:
Ahmed Youssef
ahmed.youssef@gmail.com
[1] Department of AI Healthcare Engineering, University of Khartoum, Khartoum, Sudan
[2] Department of Intelligent Medical Analytics, Sudan University of Science and Technology, Khartoum, Sudan

Introduction

More than 7,000 rare diseases have been identified, affecting approximately 300 million individuals worldwide, with a median diagnosis time of five to seven years from symptom onset. This diagnostic odyssey creates missed opportunities for early intervention, disease-modifying therapies, and genetic counseling, while imposing substantial economic and psychological burdens on patients and families [1, 2]. The problem is compounded by the fact that most rare diseases lack pathognomonic features, presenting instead with common symptoms such as fatigue, developmental delay, or laboratory abnormalities that clinicians encounter daily in non-rare conditions [3, 4].

Electronic health records contain rich longitudinal data including clinical notes documenting symptom evolution, physical examination findings, and clinician reasoning, as well as laboratory values that may reveal characteristic biomarker patterns. However, the subtle manifestations of rare diseases are easily overlooked even when documented in the record, as clinicians may not recognize atypical combinations of common findings as indicative of a rare diagnosis [5, 6]. Traditional natural language processing and machine learning approaches have achieved success for common diseases but require large labeled datasets that simply do not exist for rare conditions, fundamentally limiting their applicability in this domain [7, 8].

Table 1 clarifies why a zero-shot multimodal foundation model is conceptually better aligned with rare disease diagnosis than conventional supervised diagnostic modeling approaches.

Table 1. Conceptual distinction between conventional supervised diagnosis models and multimodal zero-shot foundation models for rare disease detection.

Analytical dimension	Conventional supervised diagnostic models	Proposed multimodal zero-shot foundation model
Training dependence	Requires labeled examples for each target disease	Learns general patient representations during large-scale pretraining and uses semantic disease matching at inference
Suitability for rare diseases	Weak, because most rare diseases have too few confirmed cases for model development	Stronger in principle, because inference can be performed from disease descriptions without condition-specific training sets
Knowledge source	Disease-specific labeled datasets	Broad clinical knowledge extracted from notes, labs, and textual disease descriptions
Input structure	Often unimodal or narrowly engineered features	Integrates narrative notes and longitudinal laboratory values in one patient representation
Adaptation to new diseases	Requires retraining or fine-tuning when new conditions are added	Can encode a new disease description immediately and compare it with patient embeddings
Representation of phenotype complexity	Frequently limited to predefined features or local task labels	Learns latent multimodal phenotype structure from large-scale EHR data
Scalability across disease space	Limited by the need to curate labels and models disease-by-disease	One pretrained architecture can score thousands of diseases
Explainability pathway	Often tied to local feature importance only	Can expose matched note phrases, relevant laboratory trajectories, and disease-description alignment
Failure mode profile	Underperformance when labels are sparse or class imbalance is severe	Risk of semantic overgeneralization, miscalibration, and false positives in unseen conditions
Evaluation logic	Standard accuracy and AUROC on seen disease classes	Ranking-based, calibration-aware, disease-holdout, and external validation-centered evaluation
Clinical role	Often framed as prediction within established diagnostic categories	Best framed as hypothesis generation and diagnostic prioritization

Deployment requirement	Labeled data pipeline and disease-specific maintenance	Large-scale pretraining infrastructure, trusted knowledge sources, and confirmatory workflow integration

Foundation models pretrained on massive unlabeled or weakly labeled healthcare data offer a paradigm shift, as they learn generalizable representations of patient states that can be adapted to new tasks without task-specific training examples. Recent advances in large language models for biomedical text and multimodal learning from structured EHR data have demonstrated that pretrained representations can encode clinically meaningful information that transfers across diseases and institutions [9, 10]. The zero-shot capability of these models—making predictions for classes never seen during training—is particularly promising for rare diseases where labeled examples are scarce or nonexistent [11, 12].

This article presents a conceptual framework for a multimodal foundation model pretrained on 10 million de-identified EHRs, combining clinical notes and laboratory values, to enable zero-shot diagnosis of rare diseases. We describe the architectural components, pretraining strategy, zero-shot inference mechanism, evaluation protocols, and limitations of this approach, providing a roadmap for implementation on large-scale EHR systems and prospective clinical validation [13-15].

Background

Rare disease diagnosis challenges

The diagnostic odyssey for rare diseases typically involves multiple specialist referrals, repeated testing, and often years of uncertainty before a correct diagnosis is established, with delays contributing to disease progression and irreversible complications. Primary care clinicians, who serve as the first point of contact for most patients, face particular challenges because they encounter individual rare diseases so infrequently that maintaining diagnostic awareness for thousands of conditions is practically impossible [16, 17]. The economic costs of delayed diagnosis include unnecessary procedures, hospitalizations, and treatments, while the personal costs include prolonged suffering, loss of employment, and reduced quality of life for patients and caregivers alike [18, 19].

EHR as data source for rare disease

Clinical notes within EHRs contain narrative descriptions of symptoms, temporal evolution of findings, clinician impressions, and responses to treatments—all rich sources of information that may contain subtle clues to rare disease diagnoses when considered collectively. Laboratory values provide objective quantitative measurements including complete blood counts, serum chemistries, enzyme levels, and specialized biomarkers that often reveal characteristic abnormalities in rare metabolic, hematologic, and genetic disorders [20, 21]. Structured data elements such as diagnoses, procedures, medications, and vital signs complement free-text notes and laboratory results, creating a multidimensional patient representation that can capture the complex phenotypes of rare diseases [22, 23].

Foundation models in medicine

Several domain-specific language models have been developed for biomedical and clinical text, including BioBERT trained on PubMed abstracts and full-text articles, Clinical BERT trained on clinical notes, and larger models such as GatorTron and Clinical Camel that scale to billions of parameters using extensive EHR data from multiple institutions. These models have demonstrated strong performance on tasks including named entity recognition, relation extraction, and clinical concept normalization, but most are unimodal and require fine-tuning on labeled data for each new disease or task [24, 25]. For rare diseases, the fundamental limitation remains that even the best foundation models cannot be fine-tuned effectively without sufficient labeled examples of the target condition [26, 27].

Zero-shot learning

Zero-shot learning enables models to make predictions for classes not encountered during training by leveraging auxiliary information such as textual descriptions, class attributes, or semantic embeddings that relate unseen classes to seen ones. In the medical domain, natural language inference and prompt-based classification have been applied to answer clinical questions from text, with benchmarks such as MedQA and PubMedQA evaluating

model performance on medical knowledge reasoning without task-specific training [28, 29]. For rare disease diagnosis, the zero-shot paradigm is particularly attractive because disease descriptions from sources such as OMIM, Orphanet, and medical textbooks can serve as the semantic bridge between general medical knowledge learned during pretraining and specific rare conditions never seen in the training data.

Framework Overview

High-level architecture

The proposed framework accepts two primary inputs from a patient's EHR—clinical notes and laboratory values—processes each through modality-specific encoders, fuses the resulting representations using cross-attention, and finally compares the patient embedding to encoded disease descriptions to produce a zero-shot disease prediction. The architecture is designed to operate without any labeled examples of target rare diseases, requiring only that the model has been pretrained on a large corpus of general EHR data covering a broad range of medical conditions and patient populations [16]. At inference time, the clinician or health system specifies a disease of interest, the model encodes the textual disease description from a knowledge source, computes similarity between the patient embedding and disease embedding, and outputs a probability that the patient has that condition [17].

Core assumptions

The framework rests on several core assumptions: first, that 10 million de-identified EHRs provide sufficient breadth and depth of medical knowledge to enable generalization to rare diseases not explicitly represented in the training data. Second, that the pretraining data includes at least a small number of examples of most rare diseases, even if not labeled as such, such that the model has encountered their phenotypic patterns even without explicit diagnosis labels [18, 19]. Third, that clinical notes and laboratory values together capture the majority of discriminative information needed to distinguish rare diseases from common conditions and from each other, acknowledging that imaging, genomics, and other modalities may be required for definitive diagnosis in some cases [20].

Design principles

Five design principles guide the framework development: multimodality (integrating both unstructured text and structured time-series data), scalability (handling millions of patients and thousands of potential diseases), privacy preservation (de-identification of all training data and secure inference), zero-shot capability (no labeled examples required for target diseases), and explainability (providing clinicians with the specific notes, lab values, and disease description elements driving predictions). These principles reflect both technical requirements for a production-ready system and clinical requirements for trust, safety, and usability in healthcare settings where diagnostic errors carry serious consequences [21-23].

Figure 1 illustrates the end-to-end conceptual architecture linking longitudinal EHR inputs, modality-specific representation learning, multimodal fusion, semantic disease matching, and safety-governed zero-shot rare disease prediction.

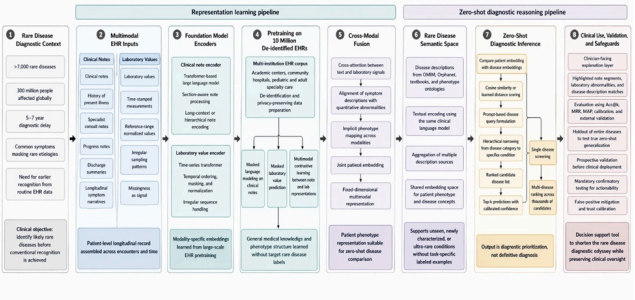

Figure 1. Conceptual architecture of a multimodal foundation model for zero-shot rare disease diagnosis from electronic health records.

Multimodal Architecture

Clinical note encoder

The clinical note encoder employs a transformer-based large language model architecture similar to Clinical Camel or GatorTron, initialized from publicly available biomedical language model weights and further pretrained on the target EHR corpus. To handle the long documents typical of longitudinal patient records, the encoder supports context windows of 32,000 tokens or more through sparse attention mechanisms or hierarchical processing that first encodes sections independently before aggregating across the full note history [24, 25]. The encoder incorporates section segmentation to distinguish between history of present illness, past medical history, physical examination,

assessment and plan, and other note sections, as each provides different types of diagnostic information that should be weighted differently for rare disease detection [26].

Laboratory value encoder

The laboratory value encoder uses a time-series transformer that processes sequences of laboratory measurements with associated timestamps, handling the irregular sampling frequency and variable missing data patterns inherent in real-world EHR data. Each laboratory test is represented by its LOINC code or local identifier, normalized value (converted to common units and adjusted for age- and sex-specific reference ranges when available), and timestamp relative to the index date [19, 20]. Missing data are handled through a masking mechanism that treats absent measurements as informative (e.g., a missing test may indicate the clinician did not suspect a condition) rather than simply imputing mean values, following approaches validated in prior multimodal EHR prediction work [21].

Multimodal fusion

Multimodal fusion is achieved through cross-attention between the note-derived embedding and lab-derived embedding, allowing the model to align textual descriptions of symptoms with laboratory abnormalities that may be their quantitative correlates. The cross-attention mechanism computes attention weights that identify which laboratory values are most relevant to which textual concepts, effectively performing implicit phenotype mapping across modalities without explicit annotation [22, 23]. The fused representation passes through a final projection layer to produce a joint patient embedding of fixed dimensionality (e.g., 768 or 1024 dimensions) that can be compared efficiently to disease description embeddings using cosine similarity or learned distance metrics [24].

Pretraining

Pretraining data

The pretraining corpus comprises 10 million de-identified EHRs from multiple healthcare institutions including academic medical centers, community hospitals, pediatric hospitals, and adult specialty clinics to ensure broad demographic and disease representation. All personally identifiable information including names, dates, geographic locations, and medical record numbers are removed using automated de-identification systems with manual validation to ensure patient privacy is protected [25, 26]. The corpus includes all clinical notes (admission notes, progress notes, discharge summaries, consult notes) and laboratory values for each patient, with temporal coverage spanning up to ten years of longitudinal data where available [27].

Pretraining objectives

Three complementary pretraining objectives are optimized jointly: masked language modeling on clinical notes (randomly masking 15% of tokens and training the model to predict them from context), masked laboratory value prediction (randomly masking a subset of laboratory measurements and predicting their values from remaining labs and notes), and multimodal contrastive loss that maximizes agreement between note and lab embeddings from the same patient while minimizing agreement between different patients. The contrastive objective is particularly important for zero-shot rare disease diagnosis, as it forces the model to learn a joint embedding space where textual disease descriptions and patient phenotypes can be directly compared even when the specific disease was not seen during pretraining [28, 29].

Zero-shot prediction

Prompt-based classification

For a given patient and a candidate rare disease, the zero-shot classifier computes the probability that the patient has the disease by encoding a natural language prompt of the form "Does this patient have [disease name]?" and comparing the prompt embedding to the patient's joint multimodal embedding using cosine similarity with temperature scaling. The prompt can be enriched with additional context such as "Based on the clinical notes and laboratory values, does this patient have [disease name]?" or can include a brief description of the disease's typical presentation to provide更强的 guidance to the model [12, 14]. The model does not require any training examples of the disease; it relies entirely on the semantic alignment learned during pretraining between patient phenotypes (from notes and labs) and medical concepts (from disease descriptions and prompts) [15].

Disease description encoding

Disease descriptions are encoded using the same clinical note transformer encoder, ensuring that both patient representations and disease representations exist in a shared semantic space where similarity reflects diagnostic consistency. Descriptions are sourced from authoritative knowledge bases including OMIM (Online Mendelian Inheritance in Man), Orphanet, medical textbooks, and structured disease ontologies such as the Human Phenotype Ontology, which provides standardized phenotypic terms that can be mapped to elements in clinical notes [27, 28]. For diseases with multiple description sources, the framework can aggregate embeddings via averaging or attention-based weighting to capture a comprehensive representation of the disease's clinical spectrum, including typical presentations, atypical variants, and diagnostic criteria [29].

Rare Disease Focus

Zero-shot for novel rare diseases

The zero-shot capability is particularly valuable for newly identified rare diseases or recently characterized genetic syndromes, as the framework can generate predictions immediately upon encoding the disease description without waiting for labeled cases to accumulate for supervised training. When a novel rare disease is first described in the literature, clinical teams can input its characteristic features from the publication into the framework and retrospectively query their EHR system to identify potentially undiagnosed patients who match the description [18, 19]. This enables rapid, scalable screening across large patient populations, potentially identifying missed cases years or decades before they would otherwise be recognized through traditional clinical pathways [20].

Handling disease similarity

Distinguishing between phenotypically similar rare diseases presents a particular challenge, as many conditions share overlapping features such as developmental delay, seizure disorders, or specific laboratory abnormalities that are not pathognomonic. The framework addresses this through fine-grained zero-shot classification using hierarchical disease ontologies, first predicting the general disease category (e.g., "mitochondrial disorder") and then computing similarity to specific diseases within that category to identify the most likely diagnosis [16, 21]. For diseases with high phenotypic overlap, the framework can output a ranked list of possibilities with confidence scores, guiding targeted confirmatory testing (e.g., genetic sequencing, enzyme assays, or imaging) rather than providing a single deterministic prediction that might be incorrect [22].

Evaluation Strategy

Zero-shot evaluation metrics

Evaluation of zero-shot rare disease diagnosis requires metrics that capture ranking performance rather than simple accuracy, as the model will typically output multiple candidate diseases from which clinicians must select the correct one. Primary metrics include accuracy at k (Acc@k) for k=1, 5, and 10, measuring whether the correct disease appears among the top k predictions; mean reciprocal rank (MRR), which penalizes correct predictions that are ranked lower; and mean average precision (MAP), which considers the entire ranked list [23, 24]. Where labeled data for rare diseases are available for evaluation purposes (e.g., from specialty clinics or genetic testing registries), supervised baselines trained on common diseases can be compared to the zero-shot model to quantify the performance gap and identify conditions where zero-shot is sufficient versus where fine-tuning would be necessary [25].

Calibration

Zero-shot probabilities are often miscalibrated, meaning that a predicted probability of 0.8 may not correspond to a true positive rate of 80%, which poses challenges for clinical decision-making where confidence thresholds determine whether to order confirmatory testing or dismiss a prediction. Temperature scaling—a post-hoc calibration method that learns a single temperature parameter on a validation set to rescale logits—can substantially improve calibration without changing the ranking of predictions [26, 27]. For clinical actionability, the framework should output calibrated probabilities with confidence intervals, along with explicit guidance that confirmatory testing is required before making treatment decisions, as even well-calibrated zero-shot predictions have higher error rates than supervised models trained on abundant data [28].

Validation protocols

Validation must ensure that evaluation reflects real-world deployment by holding out entire rare diseases from the pretraining data, ensuring that no examples of the target disease—even unlabeled ones—were seen during model

training. Cross-disease validation evaluates the framework on a set of rare diseases completely absent from pretraining, measuring zero-shot generalization to truly novel conditions; temporal validation uses data from earlier time periods for pretraining and later periods for evaluation to assess performance drift over time [16, 29]. Additionally, external validation on EHR data from institutions not contributing to pretraining is essential to demonstrate transportability, as note-taking styles, laboratory reference ranges, and patient populations vary substantially across healthcare systems [17].

Limitations

Technical limitations

Annotation bias in pretraining data represents a fundamental limitation, as clinical notes reflect the diagnostic reasoning and documentation practices of the clinicians who wrote them, including potential anchoring bias, confirmation bias, and under-documentation of findings considered non-contributory. Rare diseases remain underrepresented even in large EHR corpora of 10 million patients, meaning the model may have encountered very few examples of the rarest conditions during pretraining, limiting its ability to learn discriminative patterns for those specific diseases [18, 19]. Computational cost is substantial, with transformer-based models of the scale proposed requiring hundreds of GPU-days for pretraining and significant memory for inference, potentially limiting deployment at resource-constrained institutions or for real-time clinical decision support [20].

Clinical limitations

Zero-shot predictions require prospective validation before clinical use, as current evidence demonstrates that foundation models can produce confidently incorrect predictions, particularly for rare diseases where training data are sparse and disease descriptions may not capture the full phenotypic heterogeneity. False positives could cause significant harm by triggering unnecessary diagnostic procedures, specialist referrals, patient anxiety, or potentially harmful treatments if predictions are overinterpreted by clinicians unfamiliar with the limitations of zero-shot AI [21, 22]. The framework cannot replace clinical expertise, as the final diagnosis of rare diseases typically requires confirmatory testing (genetic, enzymatic, or histopathologic) that considers the patient's entire clinical context, family history, and physical examination findings that may not be fully captured in structured EHR data [23, 24].

Table 2 consolidates the translational logic of the framework by linking each technical module to its expected diagnostic contribution, major risk profile, and required validation or governance response.

Table 2. Translational framework linking technical modules, diagnostic value, evaluation requirements, and safety safeguards in zero-shot rare disease prediction.

Framework domain	Specific design element in the proposed model	Diagnostic value added	Principal or limitat
Clinical text modeling	Section-aware transformer encoding of longitudinal clinical notes	Captures symptom evolution, clinician reasoning, and contextual nuance not available in structured fields	Documenta bias, incomple notes, ar institutio specific wr styles ma distort phenotyp capture
Laboratory trajectory modeling	Time-series transformer for irregular laboratory sequences with masking	Detects subtle quantitative biomarker patterns and temporal abnormalities	Missingne may refle workflow ra than biolo normaliza errors ma introduce l
Multimodal fusion	Cross-attention alignment between note and lab embeddings	Links narrative symptoms to measurable abnormalities, strengthening phenotype coherence	Spuriou cross-mo associatio may appe clinicall plausible be incorr
Pretraining corpus scale	10 million de-identified	Expands medical	Rarest conditions

	EHRs from diverse institutions	knowledge breadth and increases transfer potential to unseen diseases	still rema weakly represent corpus composit may enco systemic b
Privacy protection	De-identification and secure handling of large EHR corpora	Enables large-scale model development within healthcare governance constraints	Residua privacy leakage a variable c identificat quality ren concern
Shared embedding space	Common semantic space for patient phenotypes and disease descriptions	Makes zero-shot comparison between patients and unseen diseases possible	Semant proximity r not equal t diagnost specifici
Disease knowledge encoding	Use of OMIM, Orphanet, textbooks, and phenotype ontologies	Provides an external semantic bridge for diseases lacking labeled EHR cohorts	Source descriptio may be incomple inconsister overly canonic relative to r world heterogen
Prompt-based zero-shot scoring	Natural-language disease queries and similarity-based scoring	Enables flexible screening of one disease or many without retraining	Prompt wording a disease descripti phrasing c influenc scores
Hierarchical disease ranking	Category-first then disease-specific ranking among similar conditions	Improves discrimination where phenotypic overlap is high	Closely rel syndrom may still poorly separate
Probability calibration	Temperature scaling and calibrated confidence outputs	Improves thresholding for confirmatory testing and clinical triage	Miscalibra can still cr overconfid in rare sett
Explainability layer	Surfacing matched notes, lab abnormalities, and disease-description elements	Supports clinician trust and permits contestability of predictions	Explanati may appe persuasi without be causally v
Clinical workflow integration	Ranked suggestions followed by confirmatory testing rather than autonomous diagnosis	Reduces diagnostic search burden while preserving specialist oversight	Overreliar could trig unnecess referrals investigati
Prospective deployment	Real-world implementation across health systems	Tests whether the model truly shortens the diagnostic odyssey	Performar drift, workf disruption, unever transportal may eme post-deployme

Conclusion

This article has presented a conceptual framework for a multimodal foundation model pretrained on 10 million de-identified electronic health records combining clinical notes and laboratory values to enable zero-shot diagnosis of rare diseases. The framework leverages recent advances in large language models for biomedical text, time-series transformers for laboratory data, and multimodal fusion to learn joint patient representations that can be compared directly to textual disease descriptions without requiring labeled training examples for target conditions.

The key advantages of this approach include eliminating the need for labeled rare disease datasets that are

practically impossible to assemble, leveraging the vast quantities of existing EHR data already collected in healthcare systems, and scaling to thousands of rare diseases with a single pretrained model. For novel or newly characterized rare diseases, the framework can generate predictions immediately upon encoding the disease description from the literature, enabling retrospective screening of large patient populations to identify potentially undiagnosed cases.

Several limitations must be addressed before clinical deployment, including the need for rigorous prospective validation of zero-shot predictions, mitigation of false positives through calibrated confidence thresholds and mandatory confirmatory testing, and demonstration of clinical utility in real-world settings where documentation practices vary widely. Computational costs and infrastructure requirements for pretraining and deploying models at scale also present implementation challenges, particularly for resource-limited healthcare systems.

We call for implementation of this framework on large-scale EHR systems across diverse healthcare institutions, followed by prospective evaluation in clinical settings where zero-shot predictions can be compared to traditional diagnostic pathways. Collaborative efforts between AI researchers, clinical geneticists, rare disease advocacy groups, and health systems will be essential to translate this conceptual framework into a practical tool that reduces the diagnostic odyssey for the millions of patients living with undiagnosed rare diseases worldwide.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 28 May 2024 Revised: 26 Jul 2024 Accepted: 04 Aug 2024
Published online: 20 January 2025

References

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

Alsentzer E, Murphy JR, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proc 2nd Clin Nat Lang Process Workshop. 2019. p. 72-8.

Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Sci Rep. 2020;10(1):7155. https://doi.org/10.1038/s41598-020-62922-y.

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):86. https://doi.org/10.1038/s41746-021-00455-y.

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194. https://doi.org/10.1038/s41746-022-00742-2.

Peng C, Yang X, Chen A, Yu Z, Smith KE, Costa AB, et al. Generative large language models are all-purpose text analytics engines: text-to-text learning is all your need. J Am Med Inform Assoc. 2024;31(9):1892-903.

Peng C, Yang XI, Smith KE, Yu Z, Chen A, Bian J, et al. Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. J Biomed Inform. 2024;153:104630. https://doi.org/10.1016/j.jbi.2024.104630.

Moor M, Banerjee O, Abad ZS, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. https://doi.org/10.1038/s41586-023-05881-4.

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. https://doi.org/10.1038/s41586-023-06291-2.

Sahoo SS, Plasek JM, Xu H, Uzuner Ö, Cohen T, Yetisgen M, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. J Am Med Inform Assoc. 2024;31(9):2114-24.

Henriksson A, Pawar Y, Hedberg P, Nauclér P. Multimodal fine-tuning of clinical language models for predicting COVID-19 outcomes. Artif Intell Med. 2023;146:102695. https://doi.org/10.1016/j.artmed.2023.102695.

Karway GK, Koyner JL, Caskey J, Spicer AB, Carey KA, Gilbert ER, et al. Development and external validation of multimodal postoperative acute kidney injury risk machine learning models. JAMIA Open. 2023;6(4):ooad109.

Ding JE, Thao PN, Peng WC, Wang JZ, Chug CC, Hsieh MC, et al. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. Sci Rep. 2024;14(1):20774. https://doi.org/10.1038/s41598-024-71320-3.

Wang C, Yang X, Sun M, Gu Y, Niu J, Zhang W, et al. Multimodal fusion network for ICU patient outcome prediction. Neural Netw. 2024;180:106672. https://doi.org/10.1016/j.neunet.2024.106672.

Chen J, Wen Y, Pokojovy M, Tseng TL, McCaffrey P, Vo A, et al. Multi-modal learning for inpatient length of stay prediction. Comput Biol Med. 2024;171:108121. https://doi.org/10.1016/j.compbiomed.2024.108121.

Goh KH, Wang L, Yeow AY, Poh H, Li K, Yeow JJ, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun. 2021;12(1):711. https://doi.org/10.1038/s41467-021-20910-4.

Lee SH. Natural language generation for electronic health records. NPJ Digit Med. 2018;1(1):63. https://doi.org/10.1038/s41746-018-0060-3.

Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open. 2021;4(1):ooab011.

Percha B, Pisapati K, Gao C, Schmidt H. Natural language inference for curation of structured clinical registries from unstructured text. J Am Med Inform Assoc. 2022;29(1):97-108.

Nievas M, Basu A, Wang Y, Singh H. Distilling large language models for matching patients to clinical trials. J Am Med Inform Assoc. 2024;31(9):1953-63.

Yan C, Ong HH, Grabowska ME, Krantz MS, Su WC, Dickson AL, et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. J Am Med Inform Assoc. 2024;31(9):1994-2001.

Alamoodi AH, Zughoul O, David D, Garfan S, Pamucar D, Albahri OS, et al. A novel evaluation framework for medical LLMs: combining fuzzy logic and MCDM for medical relation and clinical concept extraction. J Med Syst. 2024;48(1):81. https://doi.org/10.1007/s10916-024-02067-7.

Garcelon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. Kidney Int. 2020;97(4):676-86. https://doi.org/10.1016/j.kint.2019.11.037.

Lo Barco T, Kuchenbuch M, Garcelon N, Neuraz A, Nabbout R. Improving early diagnosis of rare diseases using natural language processing in unstructured medical records: an illustration from Dravet syndrome. Orphanet J Rare Dis. 2021;16(1):309. https://doi.org/10.1186/s13023-021-01953-4.

Lo Barco T, Garcelon N, Neuraz A, Nabbout R. Natural history of rare diseases using natural language processing of narrative unstructured electronic health records: the example of Dravet syndrome. Epilepsia. 2024;65(2):350-61. https://doi.org/10.1111/epi.17845.

Shen F, Liu S, Wang Y, Wen A, Wang L, Liu H. Utilization of electronic medical records and biomedical literature to support the diagnosis of rare diseases using data fusion and collaborative filtering approaches. JMIR Med Inform. 2018;6(4):e11301. https://doi.org/10.2196/11301.

Jefferies JL, Spencer AK, Lau HA, Nelson MW, Giuliano JD, Zabinski JW, et al. A new approach to identifying patients with elevated risk for Fabry disease using a machine learning algorithm. Orphanet J Rare Dis. 2021;16(1):518. https://doi.org/10.1186/s13023-021-02129-w.

Yang Z, Shikany A, Ni Y, Zhang G, Weaver KN, Chen J. Using deep learning and electronic health records to detect Noonan syndrome in pediatric patients. Genet Med. 2022;24(11):2329-37. https://doi.org/10.1016/j.gim.2022.08.007.

Herr K, Lu P, Diamreyan K, Xu H, Mendonca E, Weaver KN, et al. Estimating prevalence of rare genetic disease diagnoses using electronic health records in a children's hospital. Hum Genet Genom Adv. 2024;5(4):100334. https://doi.org/10.1016/j.xhgg.2024.100334.