

Symptom-Based Disease Prediction Using a T5 Transformer Model

Aanvi Manda¹

Received March 12, 2026

Accepted July 28, 2026

Electronic access November 15, 2026

Correctly diagnosing diseases from patient-reported symptoms is challenging due to vague and overlapping symptom descriptions. In this study, we evaluated whether a fine-tuned T5-small transformer model could accurately predict diseases from natural-language symptom descriptions. We hypothesized that training a lightweight sequence-to-sequence model on a structured symptom–disease dataset reformatted into text-based prompts would achieve high diagnostic accuracy. The model was trained for two epochs on a curated dataset containing 41 disease classes and evaluated on a held-out test set. This fine-tuned model achieved 99.8% exact-match accuracy. These results suggest that transformer-based models can effectively learn symptom–disease mappings, but the experimental design underscores the importance of richer clinical features, dataset quality, model interpretability, and external validation before real-world deployment.

Keywords: T5 Transformer Model, Disease Prediction, NLP

Introduction

A painfully intractable question in medical practice concerns how well one can diagnose a disease based on the symptoms as reported by a patient, as they tend to be sketchy, often inconsistent, and detailed in unstructured natural language^{1,2}. Irrespective of whether the process is executed by medical practitioners themselves or by computers, this makes the task of mapping the natural-language description of symptoms to underlying diseases arduous, and may also underlie cases of misdiagnosis and over-investigation^{3,4}. The problem is further exacerbated when diseases have more than one symptom, as the number of possible errors is then increased, and the very nature of the problem can constitute a health hazard to patients⁵. With the rise of telemedicine and online self-disease checkers, the need to develop reliable and robust tools that can make sense of patient-reported symptom descriptions has become an even more pressing concern^{6,7}.

Recent progress in transformer-based language models has made them strong candidates for symptom-driven diagnostic prognosis. Singhal et al. demonstrated that large language models are capable of acquiring a significant portion of clinical knowledge and have good performance on medical diagnostic reasoning tasks, which would suggest their usefulness in the context of clinical decision-making⁸. Indeed, they can be used to train associations between symptom description and disease diagnosis and to generate lists of possible diagnoses in a way that could help clinicians rather than replace their decision-making skills. Previous papers have also shown how the Transformer architecture could be beneficial

in medical diagnosis. Hassan et al. compared various language models to classify diseases of patients based on symptom descriptions and reported good performances for several transformer-based models, among which MCN-BERT⁹. Chen et al. proposed the use of DxFormer, an architecture based on transformer-based language model that relies on a dense representation of the symptoms to decouple the task of symptom inquiry from the task of symptom diagnosis¹⁰. Wang et al. trained a generative model, called CoAD, designed to generate symptoms and to make predictions about disease diagnosis simultaneously, in order to reinforce the diagnostic reasoning process¹¹. Tomar et al. integrated structured medical knowledge within the framework of a Transformer-based language model by building a symptom–disease knowledge graph and obtained an improvement in the task of disease identification, while also enhancing the interpretability and the clinical usefulness of the model¹². In a more general review about the use of large language models in clinical diagnosis, Zhang et al. obtained promising results using transformer-based models, and highlighted the importance of making the decisions explainable to patients and validating these models on real-world clinical data¹³.

Modern medical natural language processing techniques are to a large extent thanks to the Transformer model presented by Vaswani et al., that put forward the use of self-attention as a compelling alternative to recurrent neural networks for sequence modeling¹⁴. Raffel et al. soon after presented T5, where a multitude of different NLP tasks can be expressed in a single text-to-text paradigm¹⁵. The use of large language models to perform multiple tasks in response to a natural language instruction was then demonstrated by

¹ Lufkin High School, Texas, US

Brown et al.¹⁶. Previous works had already shown how machine learning and NLP techniques can be used to identify disease phenotypes and retrieve clinically-relevant information from patient electronic records, highlighting the growing importance of machine learning techniques for diagnostic assistance in clinical settings¹⁷. ClinicalBERT then further demonstrated how a model trained on clinical text data can be taught to acquire representations that could be employed for healthcare-related predictive tasks¹⁸. Deep learning techniques have also been used successfully to tackle the task of predicting medical codes and associated classification tasks, which again is evidence that diseases might be classified using textual data¹⁹. More recently, the T5 architecture has been modified to deal with biomedical natural language processing tasks; ClinicalT5, for example, showed that a domain-specific pretraining could lead to performance gains for a variety of clinical language-processing tasks²⁰.

A more general review of the literature on the use of artificial intelligence methods in medicine highlights that it is generally presented as a way of augmenting clinical judgement and enhancing the delivery of health services while retaining physician supervision²¹. For instance, Liu et al. found that deep learning models achieved diagnostic performance comparable to healthcare professionals in detecting diseases from medical images, although limited external validation and weaknesses in study reporting constrained the interpretation of these findings²². Other researchers have pointed out, however, that large language models developed for use in clinical settings need to be carefully assessed using rigorous evaluation techniques, human oversight, and thorough performance metrics²³. A systematic review by Busch et al. highlighted the potential of large language models in patient care while identifying limitations in reliability and explainability and emphasizing the need for expert supervision and validation of model outputs²⁴. Cho et al. provided a more expansive overview of the field, tracking the growing use of Transformer-based language models in the context of disease prediction, clinical decision support, and other medical healthcare-related tasks²⁵.

Despite these advancements, a majority of leading approaches depend on large models or elaborate, multi-stage pipelines. Thus, the performance of lightweight generative architectures in resource-constrained environments remains relatively unexplored. It is especially uncertain whether a comparatively small sequence-to-sequence model such as T5-small can achieve high diagnostic accuracy when trained on structured symptom–disease pairs recast as natural-language prompts. By investigating this question, we may contribute to the development of more accessible diagnostic tools for educational settings, low-power computing environments, and smaller clinical practices.

We hypothesized that fine-tuning a T5-small sequence-to-sequence model on a structured symptom–disease dataset

reformatted into natural-language input–output pairs would achieve high diagnostic performance, with test-set exact-match accuracy exceeding 90% in predicting disease labels from symptom descriptions.

Methods

Dataset Preparation

The dataset was retrieved from Kaggle, consisting of 4921 rows and 133 columns, with each column referring to a specific symptom out of a total of 132 possible symptoms, and each row referring to a certain classification out of a potential 41 diseases, save for the first row, which provides labels for all the symptoms. The dataset is titled “Health Symptoms and Disease Prediction Dataset”. No disease classes were removed during preprocessing. The dataset was updated recently, the old version of this dataset was used for this project. The per-class counts are listed below in the output from the console (See Figure 1)

Additionally, there are 304 unique symptom combinations.

Console output showing the number of samples associated with each of the 41 disease classes prior to train–test splitting. The dataset contained equal representation across all disease categories, with each class contributing the same number of records. This distribution was used to verify class balance before model training and evaluation.

The model selected was a T5-small parameter-efficient sequence-to-sequence transformer consisting of ~60M parameters. T5 was chosen over larger models due to its desirable balance between parameter size and state-of-the-art text to text generation power. As T5 uses a generative formulation, the disease prediction task was framed as a conditional generation problem where the input sequence was the patient’s symptoms and the output sequence was the patient’s corresponding disease. However, in order to maximize model performance, the dataset had to be reformatted to conform with the input-output pairs for a “text-in, text-out” T5 formulation. Each sample was formatted as a prompt template (“The symptoms are...”). The Kaggle data does not include free-form clinical narratives; however, it does provide binary symptom values. In order to use sequence-to-sequence training with T5, each symptom’s binary value had to be converted to a natural language statement (i.e. “The symptoms are fatigue, yellowish_skin, nausea”). However, this study thus evaluates disease prediction not from real clinical language as written by doctors, but rather from curated disease-prediction symptom-disease data in text form. Regardless, disease labels may have varying spellings or small variations within the text itself so the model was not guaranteed to generate exactly the disease name, creating challenges in evaluation with direct string matching.

Despite its popularity as a benchmark dataset for disease

target_text	
Fungal infection	120
Allergy	120
GERD	120
Chronic cholestasis	120
Drug Reaction	120
Peptic ulcer disease	120
AIDS	120
Diabetes	120
Gastroenteritis	120
Bronchial Asthma	120
Hypertension	120
Migraine	120
Cervical spondylosis	120
Paralysis (brain hemorrhage)	120
Jaundice	120
Malaria	120
Chicken pox	120
Dengue	120
Typhoid	120
hepatitis A	120
Hepatitis B	120
Hepatitis C	120
Hepatitis D	120
Hepatitis E	120
Alcoholic hepatitis	120
Tuberculosis	120
Common Cold	120
Pneumonia	120
Dimorphic hemorrhoids (piles)	120
Heart attack	120
Varicose veins	120
Hypothyroidism	120
Hyperthyroidism	120
Hypoglycemia	120
Osteoarthritis	120
Arthritis	120
(vertigo) Paroxysmal Positional Vertigo	120
Acne	120
Urinary tract infection	120
Psoriasis	120
Impetigo	120

Figure. 1 Distribution of disease classes in the Health Symptoms and Disease Prediction Dataset.

prediction, the data set lacks evidence in its documentation that the data derives from actual clinical encounters. Accordingly, the findings in this study should be interpreted as assaying disease prediction from curated symptom-disease associations rather than as from real clinical language. The data was split in training and testing sets to test model generalization; model training used the subset of the data and the held-out subset was only used for testing. In order to avoid 0-shot prediction on diseases only seen in the test set, the training data was made to be very similar to the test data in terms of disease labels in the test set. This was achieved by setting the random-state parameter for data partitioning to be 42. The final training-testing split ended up being 80% training and 20% testing (See Figure 2). No separate validation set was used. In future studies, it is important to reserve a validation partition for hyperparameter training and early stopping.

Bar chart showing the number of samples per disease used in the training (blue) and testing (orange) subsets. The dataset

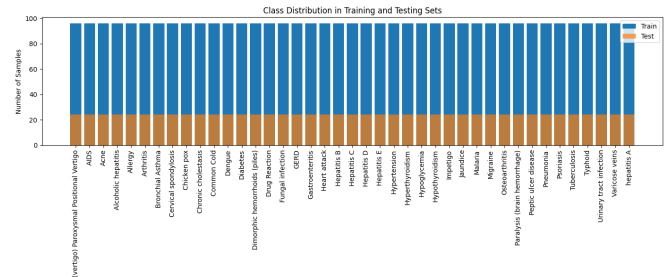


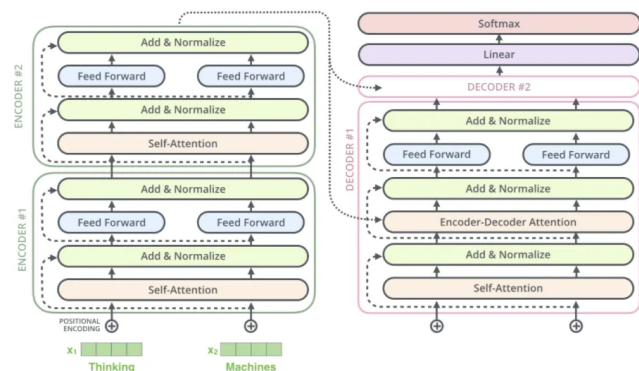
Figure. 2 Distribution of training and testing samples across disease classes.

was split using an 80:20 train-test ratio with a fixed random state of 42 to ensure consistent class representation across both subsets. All disease classes included samples in both partitions, preventing zero-shot prediction scenarios.

```
train_df, test_df = train_test_split(
    df,
    test_size=0.2,
    random_state=42,
    stratify=df["target_text"]
    # keeps label distribution similar in
    # both splits
)
```

Model Architecture

The model selected for fine-tuning was T5-small, a parameter-efficient encoder-decoder transformer containing approximately 60 million parameters. This model was chosen to balance predictive capability with computational accessibility in a Google Colab environment¹². Local development was performed on a 2020 MacBook Pro with a 1.2 GHz Quad-Core Intel Core i7 processor and 16 GB RAM.



Training Procedure

The datasets were tokenized using T5TokenizerFast with the following parameters: input max_length = 128, output max_length = 16, padding = “max_length” and truncation = True. These parameters were used to allow for standardized input and output lengths that can be trained in a batch, but also prevent a too long input for the symptoms. The tokenized datasets were then transformed into Dataset objects as is done on HuggingFace and mapped to include the encoder inputs and the decoder labels. The model was fine tuned using the API for the sequence to sequence trainer on HuggingFace, with the following hyper parameters: learning rate: 3×10^{-4} , batch size: 8 per device, number of epochs: 2, weight decay: 0.01, logging interval: 50 steps, and maximum model checkpoints stored: 2. All the models were run with the T5-small model with “conditional generation” loaded and gradient was updated only using the training subset of the data.

Following fine tuning, a prediction for each sample was obtained by running the model with each input sample in the test subset, using Hugging Face’s generate() function, with the parameters: num_beams=4, max_length=20, early_stopping=True, length_penalty=1.0. A helper function was made that took as an input the model prediction to then decode into plain text and eliminate the special tokens. Due to the nature that T5 produces free form text and not fix categorical classes, evaluating correct predictions is difficult. A simple string matching criterion was implemented as first pass evaluation: the generated predictions were lower-cased, stripped, and then exactly compared against the true diagnosis. The predictions were assessed by using: exact-match accuracy, which is a proportion of test samples that generated exactly the gold label. The disease labels were standardized before training. Exact-match accuracy is selected because it is the easiest to evaluate given all of the disease labels. Albeit the model was implemented using a text-generation framework, the task is functionally a closed-set disease classification problem since all outputs match to one of 41 labels.

Results

This experiment evaluated whether a fine-tuned T5-small sequence-to-sequence model could accurately generate disease predictions from natural-language symptom descriptions. Because the dataset consisted of structured symptom–disease pairs, testing the model on unseen symptom descriptions provided a direct way to determine how well the model generalized to new inputs.

The model was trained via the procedure discussed in Methods. This model reached a 99.8% testing accuracy, failing 2 of 984 test cases, where it predicted Chronic Cholestasis instead of Hepatitis C. A later run on this exact same model, with none

of the arguments changed, resulted in a 100% testing accuracy. Per-class precision, recall, and F1 scores were computed using this second checkpoint.. These metrics are presented to characterize class-level behavior and were not used to determine the primary accuracy reported in this study (See Table 1)

Table 1 Precision, Recall, and F1-scores by class and for the testing set

Precision, recall, F1-score, and support values computed for each of the 41 disease classes using the held-out test set. Macro-averaged and weighted-averaged metrics are reported at the bottom of the table to summarize overall model performance across all disease categories. These metrics provide a more detailed evaluation of classification performance than exact-match accuracy alone by quantifying the model’s ability to correctly identify each disease class.

	precision	recall	f1-score	support
(vertigo) Paroxysmal Positional Vertigo	1.0000	1.0000	1.0000	24
AIDS	1.0000	1.0000	1.0000	24
Acne	1.0000	1.0000	1.0000	24
Alcoholic hepatitis	1.0000	1.0000	1.0000	24
Allergy	1.0000	1.0000	1.0000	24
Arthritis	1.0000	1.0000	1.0000	24
Bronchial Asthma	1.0000	1.0000	1.0000	24
Cervical spondylosis	1.0000	1.0000	1.0000	24
Chicken pox	1.0000	1.0000	1.0000	24
Chronic cholestasis	1.0000	1.0000	1.0000	24
Common Cold	1.0000	1.0000	1.0000	24
Dengue	1.0000	1.0000	1.0000	24
Diabetes	1.0000	1.0000	1.0000	24
Dimorphic hemorrhoids (piles)	1.0000	1.0000	1.0000	24
Drug Reaction	1.0000	1.0000	1.0000	24
Fungal infection	1.0000	1.0000	1.0000	24
GERD	1.0000	1.0000	1.0000	24
Gastroenteritis	1.0000	1.0000	1.0000	24
Heart attack	1.0000	1.0000	1.0000	24
Hepatitis B	1.0000	1.0000	1.0000	24
Hepatitis C	1.0000	1.0000	1.0000	24
Hepatitis D	1.0000	1.0000	1.0000	24
Hepatitis E	1.0000	1.0000	1.0000	24
Hypertension	1.0000	1.0000	1.0000	24
Hyperthyroidism	1.0000	1.0000	1.0000	24
Hypoglycemia	1.0000	1.0000	1.0000	24
Hypothyroidism	1.0000	1.0000	1.0000	24
Impetigo	1.0000	1.0000	1.0000	24
Jaundice	1.0000	1.0000	1.0000	24
Malaria	1.0000	1.0000	1.0000	24
Migraine	1.0000	1.0000	1.0000	24
Osteoarthritis	1.0000	1.0000	1.0000	24
Paralysis (brain hemorrhage)	1.0000	1.0000	1.0000	24
Peptic ulcer disease	1.0000	1.0000	1.0000	24
Pneumonia	1.0000	1.0000	1.0000	24
Psoriasis	1.0000	1.0000	1.0000	24
Tuberculosis	1.0000	1.0000	1.0000	24
Typhoid	1.0000	1.0000	1.0000	24
Urinary tract infection	1.0000	1.0000	1.0000	24
Varicose veins	1.0000	1.0000	1.0000	24
hepatitis A	1.0000	1.0000	1.0000	24
accuracy			1.0000	984
macro avg	1.0000	1.0000	1.0000	984
weighted avg	1.0000	1.0000	1.0000	984

To test robustness outside the test set used, further manual testing was performed using symptom combinations that were manually entered in the console. Symptoms were put in different orders in the input sequence, without changing the predicted diagnosis. This suggested that the model was relatively robust to the order of the symptoms and focused more on the content of the symptoms rather than its position in the input sequence. These findings were qualitative and therefore not included in the qualitative metrics.

Evaluation of the fine-tuned T5-small model demonstrated

near-perfect performance across all metrics. The model achieved an evaluation loss of 5.2×10^{-5} and an exact-match accuracy of 1.0, indicating that every generated disease label matched the corresponding ground-truth label in the test set. Text-based evaluation metrics further supported this outcome: ROUGE-1, ROUGE-L, and ROUGE-Lsum scores were all 1.0, reflecting complete token-level and sequence-level overlap between predicted and reference outputs. The ROUGE-2 score (0.487) was lower, which is expected for extremely short outputs such as one- or two-word disease names, where bigram overlap provides limited discriminative power (See Table 2).

Table 2 Evaluation metrics for the fine-tuned T5-small model. Exact-match accuracy, ROUGE scores, evaluation loss, throughput, and runtime are reported for the held-out test set. Metrics were computed by comparing generated disease labels against ground-truth labels using exact string matching and token-level overlap measures.

Metric	Value
Exact-match accuracy	1.00
ROUGE-1	1.00
ROUGE-2	0.487
ROUGE-L	1.00
ROUGE-Lsum	1.00
Evaluation loss	0.000052
Samples/second	82.32
Runtime (seconds)	11.95

Discussion

The T5 model had strong performance on the symptom-based disease prediction task, achieving a 99.8% testing accuracy after two epochs of fine-tuning. This high level of accuracy suggests that the model effectively learned patterns within the dataset and that it can successfully apply these patterns to unseen examples. The two misclassifications seen in the results indicate the challenges inherent in predicting a diagnosis when two diseases share near-identical symptom profiles. The input symptoms for those two errors (fatigue, yellowish_skin, nausea, loss_of_appetite, and yellowing_of_eyes) were common between these two diseases, thus leaving little information with which to distinguish them. In contrast, symptoms such as itching, vomiting, and abdominal pain are found only in the case of Chronic Cholestasis within this dataset, and fatigue and family_history are only present in Hepatitis C. When these differentiating features are included, the model classifies the cases correctly. These errors highlight a more general issue not of model weakness but rather of dataset ambiguity. In

actual clinical practice, Hepatitis C and cholestatic disorders often require lab testing, imaging, or a detailed medical history in order for a correct diagnosis to be made because initial presentations may be clinically indistinguishable. The model’s confusion in such cases of ambiguity is thus more reflective of the ambiguity present within the dataset itself rather than a failure of the model to learn the symptom-disease relationships.

At the same time, the results highlight important limitations in symptom-only diagnostic modeling. The dataset is unable to capture the severity or time course of a symptom, comorbidities, biochemical markers, and other factors that strongly influence clinical decision-making. Without such additional features, some degree of ambiguity is inherent and unavoidable for some diagnoses. In addition, the distribution of symptoms across diseases may introduce implicit biases in the model. Diseases that appear with more diverse or distinctive combinations of symptom values will be easier to identify, while diseases that overlap in their presentations will be inherently more difficult to distinguish. Any automated disease-prediction model risks being inappropriately interpreted as medical advice. Incorrect predictions may contribute to delayed treatment, inappropriate self-diagnosis, or needless anxiety. The model presented here, then, should be used as a research tool but not as a stand-alone clinical decision-support system.

Additionally, the experiment would have benefited from further replicates and comparative controls. Training the model with multiple random train-test splits or evaluating the model at multiple epochs could help determine whether or not the high accuracy seen here is stable or depends on a particular split. Inclusion of baseline models such as BERT or T5-big would provide an even clearer benchmark for evaluating relative T5-small performance.

This study used a single stratified 80/20 train-test split and did not perform repeated random splits or k-fold cross validation. As a result, the reported accuracy may be sensitive to the particular partition. In future work, repeated validation procedures and explicit duplicate-combination analyses should be included in order to better estimate model robustness. Another limitation is the lack of baseline comparisons. Traditional machine-learning methods such as logistic regression, random forests, XGBoost, and nearest-neighbor classifiers were not assessed here. As the dataset consists largely of structured symptom indicators, there is value in comparing T5-small to these simple methods to see if transformer-based generation offers measurable gains. T5-small was chosen because it offers significantly lower computational demands while still achieving very high predictive accuracy. T5-small with its ~60M parameters tends to require less memory, less training time, and fewer overall computational resources than the larger models, making it more accessible for use by

researchers in resource-constrained environments. Given the high 99.8% accuracy achieved in this study, it is not clear that the additional computation of larger models would necessarily provide meaningful improvements in practical performance on this dataset. However, comparison to larger T5 models can ascertain whether the increased model capacity also provides measurable gains in diagnostic performance.

Future work should also test the inclusion of richer clinical features – such as laboratory values, patient demographics, or temporal patterns of symptoms – in order to decrease ambiguity and increase diagnostic precision. Alternative training strategies such as contrastive learning or class-balanced sampling may also assist the model in better discriminating diseases that overlap in presentations. Further, evaluation of the model on external datasets or on real-world clinical notes would provide a useful insight into model robustness outside the curated training environment.

All in all, the results show that transformer-based language models such as T5 are capable of high accuracy in symptom-based prediction tasks, particularly when there are sufficient discriminative symptoms available. The few errors seen here highlight not a failing of the model but the inherent limitations of symptom-only diagnostic data and point to promising avenues for future work.

References

- 1 C. Dreisbach, T. A. Koleck, P. E. Bourne, S. Bakken. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*. Vol. 125, pg. 37–46, 2019. <https://doi.org/10.1016/j.ijmedinf.2019.02.008>.
- 2 S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*. Vol. 17, pg. 128–144, 2008. <https://doi.org/10.1055/s-0038-1638592>.
- 3 H. Singh, G. D. Schiff, M. L. Graber, I. Onakpoya, M. J. Thompson. The global burden of diagnostic errors in primary care. *BMJ Quality & Safety*. Vol. 26, pg. 484–494, 2017. <https://doi.org/10.1136/bmjqs-2016-005401>.
- 4 M. L. Graber. The incidence of diagnostic error in medicine. *BMJ Quality & Safety*. Vol. 22, Suppl. 2, pg. ii21–ii27, 2013. <https://doi.org/10.1136/bmjqs-2012-001615>.
- 5 O. Kostopoulou, B. C. Delaney, C. W. Munro. Diagnostic difficulty and error in primary care – a systematic review. *Family Practice*. Vol. 25, pg. 400–413, 2008. <https://doi.org/10.1093/fampra/cmn071>.
- 6 H. L. Semigran, J. A. Linder, C. Gidengil, A. Mehrotra. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*. Vol. 351, Article h3480, 2015. <https://doi.org/10.1136/bmj.h3480>.
- 7 D. Chambers, J. Cantrell, M. Johnson, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open*. Vol. 9, Article e027743, 2019. <https://doi.org/10.1136/bmjopen-2018-027743>.
- 8 K. Singhal, S. Azizi, T. Tu, et al. Large Language Models Encode Clinical Knowledge. *Nature*. Vol. 620, pg. 172–180, 2023.
- 9 E. Hassan, T. Abd El-Hafeez, M. Y. Shams. Optimizing Classification of Diseases Through Language Model Analysis of Symptoms. *Scientific Reports*. Vol. 14, Article 1507, 2024. <https://doi.org/10.1038/s41598-024-51615-5>.
- 10 W. Chen, C. Zhong, J. Peng, Z. Wei. DxFormer: A Decoupled Automatic Diagnostic System Based on Decoder–Encoder Transformer with Dense Symptom Representations. *Bioinformatics*. Vol. 39, Article btac744, 2023.
- 11 H. Wang, W. C. Kwan, K.-F. Wong, Y. Zheng. CoAD: Automatic Diagnosis Through Symptom and Disease Collaborative Generation. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Vol. 1: Long Papers. pg. 6348–6361, 2023. <https://doi.org/10.18653/v1/2023.acl-long.350>.
- 12 M. Tomar, A. Tiwari, S. Saha. Towards knowledge-infused automated disease diagnosis assistant. *Scientific Reports*. Vol. 14, Article 13442, 2024. <https://doi.org/10.1038/s41598-024-53042-y>.
- 13 X. Zhang, Y. Wang, H. Li, et al. Large Language Models for Clinical Diagnosis: A Systematic Review. *Journal of Biomedical Informatics*. Vol. 157, Article 104668, 2024.
- 14 A. Vaswani, N. Shazeer, N. Parmar, et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*. Vol. 30, 2017.
- 15 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*. Vol. 21, pg. 1–67, 2020. <https://jmlr.org/papers/v21/20-074.html>.
- 16 T. B. Brown, B. Mann, N. Ryder, et al. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*. Vol. 33, pg. 1877–1901, 2020.
- 17 C. Shivade, P. Raghavan, E. Fosler-Lussier, et al. A Review of Approaches to Identifying Patient Phenotype Cohorts using Electronic Health Records. *Journal of the American Medical Informatics Association*. Vol. 21, pg. 221–230, 2014.
- 18 K. Huang, J. Altsaar, R. Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *Proceedings of Machine Learning Research*. Vol. 106, pg. 72–89, 2019.
- 19 J. Mullenbach, S. Wiegrefe, J. Duke, et al. Explainable Prediction of Medical Codes from Clinical Text. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. pg. 1101–1111, 2018.
- 20 Q. Lu, D. Dou, T. Nguyen. ClinicalT5: A Generative Language Model for Clinical Text. *Findings of the Association for Computational Linguistics: EMNLP 2022*. pg. 5436–5443, 2022. <https://doi.org/10.18653/v1/2022.findings-emnlp.398>.
- 21 E. Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.
- 22 X. Liu, L. Faes, A. U. Kale, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*. Vol. 1, pg. e271–e297, 2019. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2).
- 23 T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, et al. A Framework for Human Evaluation of Large Language Models in Healthcare Derived from Literature Review. *npj Digital Medicine*. Vol. 7, Article 258, 2024. <https://doi.org/10.1038/s41746-024-01258-7>.
- 24 F. Busch, L. Hoffmann, C. Rueger, et al. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*. Vol. 5, Article 26, 2025. <https://doi.org/10.1038/s43856-024-00717-2>.
- 25 H. N. Cho, T. J. Jun, Y. H. Kim, et al. Task-Specific Transformer-Based Language Models in Health Care: Scoping Review. *JMIR Medical Informatics*. Vol. 12, Article e49724, 2024.
- 26 J. Alammam. *The illustrated transformer*. <https://jalammar.github.io/illustrated-transformer/>. 2025.