DRGCoder: Explainable Clinical Coding for the Early Prediction of Diagnostic-Related Groups

Daniel Hajialigol[1], Derek Kaknes[2], Tanner Barbour[2], Daphne Yao[1], Chris North[1], Jimeng Sun[3], David Liem[4], Xuan Wang[1]
[1]Department of Computer Science, [2]Department of Statistics, Virginia Tech
[3]Department of Computer Science, University of Illinois Urbana-Champaign
[4]Department of Medicine, University of California, Davis
[1,2]{danielhajialigol,derekkaknes,tannerb3,danfeng,north,xuanw}@vt.edu,
[3]{jimeng}@illinois.edu, [4]{daliem}@ucdavis.edu

Abstract

Medical claim coding is the process of transforming medical records, usually presented as free texts written by clinicians, or discharge summaries, into structured codes in a classification system such as ICD-10 (International Classification of Diseases, Tenth Revision) or DRG (Diagnosis-Related Group) codes. This process is essential for medical billing and transitional care; however, manual coding is time-consuming, error-prone, and expensive. To solve these issues, we propose DRGCoder[1,2], an explainability-enhanced clinical claim coding system for the early prediction of medical severity DRGs (MS-DRGs), a classification system that categorizes patients' hospital stays into various DRG groups based on the severity of illness and mortality risk. The DRGCoder framework introduces a novel multi-task Transformer model for MS-DRG prediction, modeling both the DRG labels of the discharge summaries and the important, or *salient* words within he discharge summaries. We allow users to inspect DRGCoder's reasoning by visualizing the weights for each word of the input. Additionally, DRGCoder allows users to identify diseases within discharge summaries and compare across multiple discharge summaries.

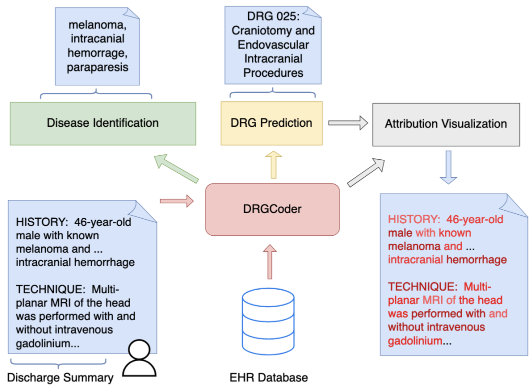


Figure 1: Overview of DRGCoder. Given a discharge summary, DRGCoder (1) identifies diseases, (2) predicts the corresponding DRG, and (3) highlights the importance of each word in the discharge summary contributing to the DRG prediction.

1 Introduction

Inpatient care, defined as care for hospital patients who stay overnight, are one of the biggest components in healthcare costs, accounting for approximately 31% of total medical expenses (Muka et al., 2015). Appropriate determination of costs associated with inpatient care are based on assigning one (of potentially many) ICD or DRG codes to a given patient. This is a crucial process for medical insurance billing and healthcare improvement, but it can be very time-consuming, prone to errors, and expensive when done manually. On average, it takes a medical coder about 20 minutes to code a single inpatient stay, and with 35 million inpatient stays in the United States each year, manual coding can be a very laborious and expensive process.

The current process for coding inpatient records most often consists of a Certified Inpatient Coder (CIC) manually reviewing a medical record and identifying every codable entity within that text. Often times, this process is completed across two completely separate software applications: the medical record is either accessed via the electronic health record (EHR) system or else "printed", usually in a digital (pdf) format, for the coder to review, while the ICD-10 codes are entered into a separate computer-assisted coding (CAC) application. These collection of ICD codes are then used to determine the correct DRG code. Alternatively, early prediction of DRG codes attempt to identify DRG codes directly from the discharge summary, bypassing ICD code prediction.

There have been many attempts to automate clinical coding while incorporating explainability. Al-

[1]Our demo is available at https://huggingface.co/spaces/danielhajialigol/DRGCoder

[2]A video demonstrating the demo can be found at https://www.youtube.com/watch?v=pcdiG6VwqlA

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 373–380
December 6-10, 2023 ©2023 Association for Computational Linguistics

though there are many methods that compute attention for each word from the input discharge summary (Dong et al., 2021; Mullenbach et al., 2018; Khalid et al., 2022; Liu et al., 2021), most of them don't utilize the powerful contextual embeddings backed by Transformer (Vaswani et al., 2017) models. Only recently however has there been Transformer-based methodologies for automating clinical coding (Wang et al., 2022; Trigueros et al., 2022). An example of this is Medical Concept Driven Attention (Wang et al., 2022), where they align both clinical notes and Wikipedia documents into topic space via topic modeling. While their architecture can handle any encoder, they claim that using a Transformer as their encoder produces inferior results, as documents are too long. Our work is most similar to Trigueros et al., where they use a multi-task Transformer-based approach, except they perform entity-linking on medical concepts related to ICD codes. Despite advancements in both Wang et al.; Trigueros et al., they don't take into account salient words: words that are *explicitly* important for clinical coding.

To address these issues, we propose DRGCODER (Figure 1), an explainable clinical coding system that provides both a predicted MS-DRG code and highlighted areas of interest within the discharge summary text to support medical coders. DRGCODER first predicts the appropriate DRG based on a novel multi-task Transformer architecture that incorporates both discharge summaries and salient words within the summaries. While there are many types of DRG codes, we choose to focus on identifying MS-DRG codes, as the MS-DRG system is the most widely used, standardized DRG system, and it is maintained by a single entity (The Centers for Medicare & Medicaid Services). DRGCODER allows users to gain better insight into MS-DRG classification by highlighting the areas of the discharge summary that explain the prediction. Furthermore, DRGCODER identifies diseases within discharge summaries and functionality to compare DRGCODER results across multiple discharge summaries.

1. We propose DRGCODER, an explainable clinical claim coding system for the early prediction of MS-DRGs. DRGCODER is supported by a novel multi-task text classification that learns to identify MS-DRG codes and important word.

2. We visualize the importance each word has on the MS-DRG prediction via a heatmap, allowing for researchers to understand DRGCODER predictions.

3. Information extraction processes, such as named-entity recognition and dense passage retrieval are incorporated to identify diseases within a discharge summary and find related discharge summaries, respectively.

2 Related Work

2.1 Early DRG Prediction

The early DRG prediction literature is shallow. Numerous machine learning techniques, such as Naive Bayes, Bayesian Networks, and Decision Trees have been applied for early DRG prediction (Gartner et al., 2015); however, they hand-crafted many of their features. In contrast, DRGCODER automatically learns features through the usage of contextualized embeddings. The framework proposed in Liu et al. applies the Convolutional Attention for Multi-Label classification (CAML) framework (Mullenbach et al., 2018) to MS-DRG codes and additionally retrieves the most important input words based on the prediction. Their framework, unlike DRGCODER, does not leverage the power of contextualized embeddings via Transformer models.

2.2 Canonnical DRG Prediction

Canonnical DRG prediction systems require the availability of ICD codes for each patient. Unfortunately, many of these systems are closed-source, as many are commercialized[3,4,5]; however, there are open-source models that have been designed to tackle this problem. DEEPDRG (Islam et al., 2021) utilizes gated recurrent units (GRUs) to predict DRG codes; however, the authors only focus on DRG codes related to urinary diseases. This significantly impacted the granularity of the classification problem, resulting in the label space consisting of only 200 DRG codes. AMANet (He et al., 2020) predicts DRGs by viewing the input data from a multi-view perspective; each input is represented by the given diagnosis and procedures of a given patient. KG-MTT-BERT (He et al., 2022) is a multi-modal model that embeds clinical notes using BERT (Devlin et al., 2018) and a custom

[3]https://nym.health/
[4]https://www.3m.com/
[5]https://www.artificialmed.com/

Table 1: Statistics for the MS-DRG-related data from MIMIC-III. Note that DRGCODER only utilizes MS-DRG patient-related data.

Patients	18,132
Hospital Stays	21,440
Unique DRGs	738

medical knowledge graph. Both KG-MTT-BERT and AMANet, however, do not provide explainable architectures, unlike DRGCODER.

2.3 Explainable Clinical Coding

Previous research has utilized label-wise attention mechanisms to highlight key elements such as n-grams (Mullenbach et al., 2018), words, and sentences (Dong et al., 2021) during the coding process. However, there is a need for further research to evaluate their effectiveness and incorporate more inherently explainable methods. Additionally, current methods primarily focus on explaining model decisions by analyzing model attention, and there have been observations of correlations between model attention and human attention (Atanasova et al., 2020; Sen et al., 2020; López-García et al., 2023; Wang et al., 2022; Trigueros et al., 2022; Khalid et al., 2022; Liu et al., 2022). Nevertheless, there have been limited research efforts to actively align model attention with human attention during the training of machine learning models.

3 System Description

DRGCODER can be broken down into two categories: automated clinical coding and explanation visualization. The former component is home to our MS-DRG classification framework while the latter contains our related summary extraction and disease named-entity recognition modules.

3.1 Automated Clinical Coding

Given a discharge summary, our automated clinical coding module identifies salient words located in discharge summaries and predicts MS-DRG codes. For MS-DRG code classification, we develop a novel multi-task BERT-based Transformer model that incorporates both discharge summaries and salient words.

Database The external database used for DRG classification is MIMIC (Johnson et al., 2018), which consists of health-related data for over 40,000 patients of the Beth Israel Deaconess Medical Center between 2001 and 2012. In addition to EHRs, MIMIC also contains billing system information, mainly comprising of ICD and DRG codes. The clinical information for each patient is captured in a discharge summary, which has an associated DRG code and (at least one) ICD code. We utilize the discharge summary and MS-DRG code across all components in our system. Although there are over 40,000 patients, only around 18,000 patients have MS-DRG data. Table 1 displays MS-DRG data statistics from MIMIC.

DRG Classification DRGCODER predicts a DRG code given a discharge summary and salient words from the discharge summary, which should play a strong role in determining the corresponding DRG code. Ideally, we would have a gold standard dataset as part of our training process: discharge summaries with salient words identified by clinical experts; however, as manual clinical coding is labor intensive, no such dataset exists, to the best of our knowledge. Thus, we chose to operate under the weakly-supervised paradigm by automatically extracting words indicative of ICD codes using BioPortal (Noy et al., 2009), a database of biomedical ontologies. Note that DRGCODER still falls under the early DRG prediction paradigm as we don't use the ICD codes identified by the CICs.

We then input both discharge summaries and ICD concepts into a novel multi-task BERT-based Transformer that learns to jointly identify DRG codes and salient words. Figure 2 demonstrates an example of this module at play. BioPortal identifies the ICD concepts "chest pain", "pain", "edema", "cough", "abdominal pain", and "diarrhea", which are then used as part of the DRG classification process. Additionally, we include external links to both identified ICD concepts and DRG codes for users to explore further.

Algorithm Design We propose an explanation-enhanced clinical coding method that aligns the model attention with human attention during training (Fig. 3). In addition to predicting the DRG code of a discharge summary, we introduce an auxiliary task that attempts to predict the salient tokens in the discharge summary. More formally, given a discharge summary $S = \{t_i | \forall i \in [1, |S|]\}$ composed of tokens t_i and corresponding saliency labels $A = \{a_i | \forall i \in [1, |S|]\}$, where $a_i = 1$ if token t_i is salient and 0 otherwise, we jointly predict

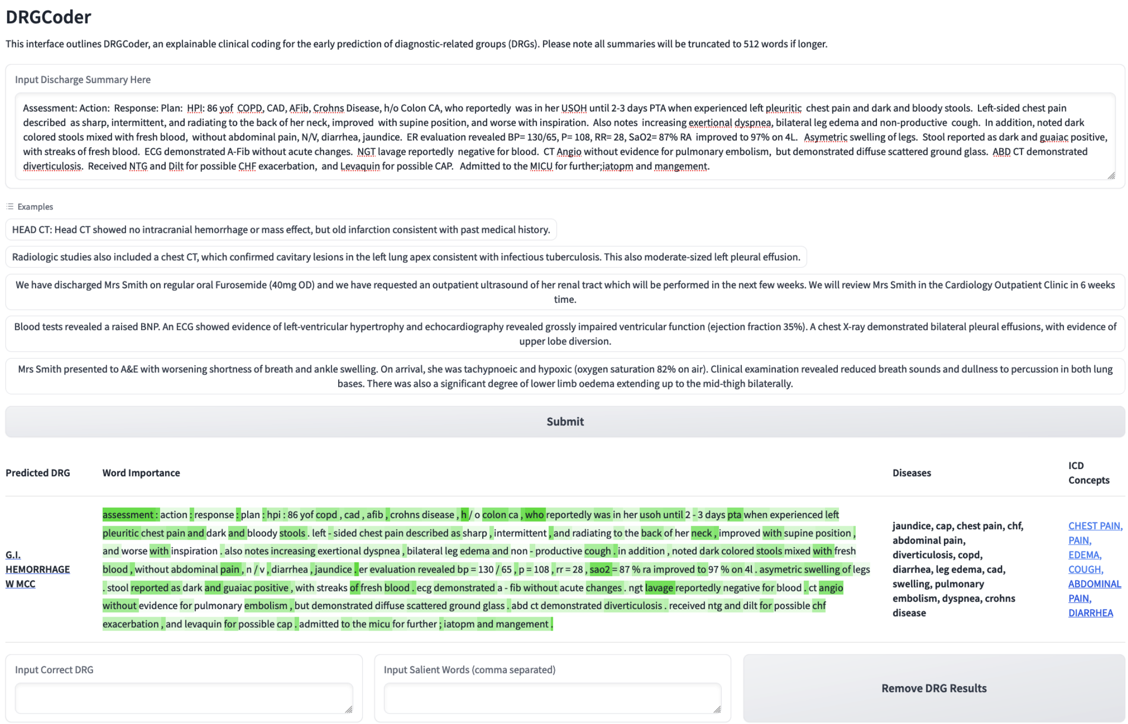


Figure 2: An example of DRGCODER using the given discharge summary. We identify the predicted DRG and ICD concepts, diseases, and attributions (the darker the highlighted token is, the more important and is, and vice-versa.). Additionally, we allow the user to input the correct DRG and important tokens, if available.

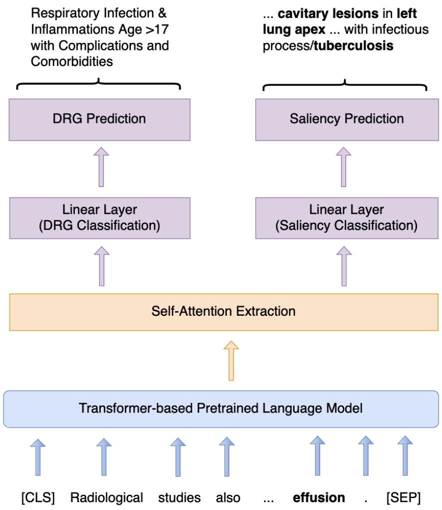


Figure 3: DRGCODER's DRG Prediction algorithm. A discharge summary and salient tokens ("effusion", in this example) are given as input. A Transformer-based language model embeds the discharge summary. For DRG prediction, these embeddings are mapped to a linear layer to get the final prediction. For saliency prediction, the self-attention matrix from the last layer and the last attention head is extracted from the Transformer model. This matrix is then mapped by a linear layer to get (binary) saliency predictions for each token.

the DRG code of the sequence as well as salient tokens.

We use a Clinical Transformer-based pre-trained language model as our backbone (Alsentzer et al., 2019), which has been pretrained on clinical data. For DRG prediction, we employ a cross-entropy loss function l_{DRG}, where the model prediction is generated by applying a linear classification layer on top of the contextualized embeddings. Saliency classification, on the other hand, utilizes a binary cross-entropy loss function l_S, with the model prediction obtained by applying a linear classification layer on the last Transformer layer and last attention head for each token. To combine both tasks, we define the final multi-task loss function as a linear combination of the aforementioned two loss functions, resulting in $l = l_{\text{DRG}} + \lambda l_S$, where λ controls the trade-off between the two tasks.

3.2 Explanation Visualization

Users are able to extract further insight into discharge summaries through our attribution visualization, related summary extraction modules. We offer functionality to visualize word importance for the most confident DRG prediction, find related discharge summaries, and extract diseases living in a queried discharge summary.

Attribution Visualization Given that misclassifying DRG codes could lead to revenue loss (Ayub et al., 2019; Zafirah et al., 2018), researchers may want to understand why a MS-DRG code was predicted for a given discharge summary. DRGCODER employs a heat-map on the weights of the DRG Classification module to visualize the most important words in the input. The darker the highlighted word is, the more important the word is. In Figure 2, for example, DRGCODER identified salient words for the input discharge summary was "assessment", "angio", "lavage", "guaiac positive", etc. DRGCODER is able to understand that the discharge summary is correctly classified as "gastrointestinal hemorrhage". This is because "angio", short for "angiogram", and lavage are both acceptable techniques in identifying gastrointestinal bleeding (Kim et al., 2014; Ousterhout and Feller, 1968). This allows for users to better understand the reasoning for the DRG prediction DRGCODER makes.

Related Summary Extraction DRGCODER supports comparison of discharge summary results (predicted DRG, word attribution, diseases, ICD concepts) across *similar* discharge summaries. The degree of similarity between a pair of discharge summaries is dependent upon BioSimCSE (raj Kanakarajan et al., 2022), which trained biomedical sentence embeddings for sentence similarity using SimCSE (Gao et al., 2021). Sentence embeddings were computed by employing contrastive learning (Eq. 1), a training framework that attempts to learn an embedding space where embeddings of similar entities, z_i and z_j, are learned to be close together, for some similarity metric (typically cosine similarity). SimCSE and BioSimCSE chose the pairing of (z_i, z_j) to correspond to the same entity, x_i, but they input x_i to a Transformer *twice* in order for the embeddings to obtain different dropout masks (as the dropout probability is random for each embedding).

$$l_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j))}{\sum_{k=1,k\neq i}^{N} \exp(\text{sim}(z_i, z_k))} \quad (1)$$

Figures 4a and 4b illustrate the related summary extraction module. In Figure 4a, users are displayed the five most similar discharge summaries to the query discharge summary (same as the example in Figure 2), with the similarity scores listed in the beginning of each discharge summary. The user can select multiple related discharge summaries to compare with the original query discharge summary, exemplified in Figure 4b. Users can compare and contrast the DRG prediction, word attribution, diseases, and ICD concepts identified across different discharge summaries. Discharge summaries that have similar diseases and/or salient words, for example, are more likely to be categorized under the same DRG code, or at the very least be classified to related DRG codes. This is also true when comparing ICD concepts, as ICD concepts are more granular than DRG codes, implying that if two summaries have similar ICD concepts, their corresponding DRG codes could be the same/related.

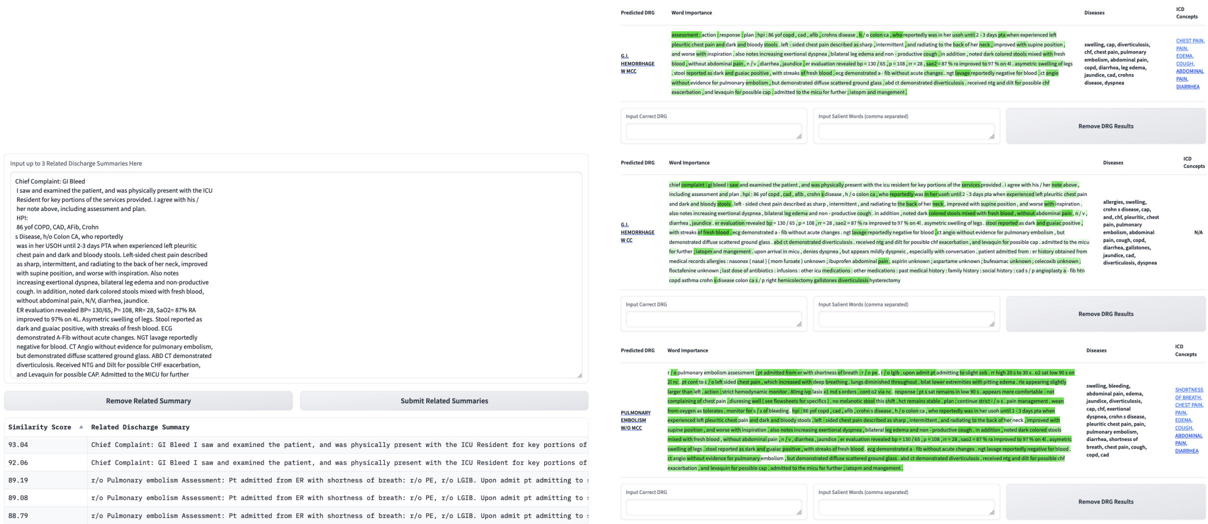
Table 2: Macro and Micro AUC and F1 scores of Clinical-BERT and CAML baselines against DRGCODER on the MIMIC-III dataset. For our experiments, we set $\lambda = 0.5$.

Model	Macro-AUC	Micro-AUC	Macro-F1	Micro-F1
CAML	0.871	0.956	0.084	0.270
Clinical-BERT	0.901	0.962	**0.106**	0.256
DRGCODER	**0.911**	**0.969**	0.101	**0.279**

Disease Named-Entity Recognition Since the MS-DRG coding system takes into account medical severity of a patient's condition, it is intuitive that diseases play a pivotal role in DRG prediction. Thus, DRGCODER provides functionality for identifying the diseases associated with a given discharge summary. This module depends on bioNLP (Alonso Casero, 2021), a system that identifies diseases from biomedical text. Figure 2 shows an example of interface for displaying these results. Given the query in Figure 2, DRGCODER lists the diseases between the "Word Importance" and "ICD Concepts" columns, namely "jaundice", "diverticulosis", "diarrhea", etc. Bleeding in the upper gastrointestinal tract has been shown to occur when patients have jaundice (Dixon et al., 1984), while diverticulosis and diarrhea are both conditions that have also been known to result in bleeding throughout the gastrointestinal tract. While this module is not part of the DRG classification process, we include this functionality to allow users to parse the discharge summary quicker so that users can validate DRG predictions faster.

4 Evaluation

We illustrate the effectiveness of DRGCODER by comparing the performance of Clinical-BERT and



(a) Related discharge summary pool.

(b) Discharge summary results comparison

Figure 4: (a) Related discharge summary pool. Users can click multiple related discharge summaries to input in the box on the top left. The similarity scores for each related discharge summary are displayed at the beginning of each related summary. (b) Results comparison across multiple discharge summaries.

CAML (Liu et al., 2021) on the MS-DRG portion of the MIMIC-III dataset. We evaluate our performance using F1 and AUC (area under curve) performance metrics. Table 2 indicates that the DRGCODER system outperforms all other compared frameworks. Specifically, when compared to Clinical-BERT, it demonstrates the effectiveness of using salient words as input, making the framework simultaneously explainable and more performant.

5 Conclusion

We offer DRGCODER, an explainability-enhanced inpatient claim coding system for MS-DRG code prediction. DRGCODER allows users to visualize words deemed important by our framework, as well as identify diseases and ICD concepts. Furthermore, we offer functionality for similar summary retrieval and comparison across summaries.

6 Future Work

We believe a natural extension of DRGCODER would be to incorporate human feedback. We ideally would like to gather feedback about correct DRGs and important words from users who use DRGCODER. Additionally, we would like to effectively incorporate the hierarchy of DRG codes. Although there aren't many levels in the DRG taxonomy, taking into account parent and sibling nodes (DRG codes) and finding appropriate information via knowledge graphs could offer better insight on the subtleties between adjacent DRG codes. For billing purposes, it would still be insightful for an incorrect DRG prediction to be classified under the same group of DRG codes, as this would indicate similar billing between the incorrectly predicted and ground truth DRG codes.

7 Limitations

A limitation of this work is the sole reliance on The MIMIC-III database. While this is a popular database, it only contains medical information from one hospital. Using only this database may not capture health-related information that occurs outside of the sampled population.

8 Ethics Statement

The usage of MIMIC requires researchers to ethical principles and guidelines that come with using electronic health records. Although MIMIC is deidentifed and open-source, it does contain sensitive and confidential patient information collected, and its use requires the utmost consideration for ethical practices. We believe our purpose falls in line with these values, as we only provide functionality for DRG prediction and analysis. The authors have been granted access to MIMIC by the MIT Laboratory for Computational Physiology.

9 Acknowledgement

Our work is sponsored by the Commonwealth Cyber Initiative and a generous gift from the Amazon + VT Center for Efficient and Robust ML.

References

Álvaro Alonso Casero. 2021. *Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature*. Ph.D. thesis, ETSI_Informatica.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. *arXiv preprint arXiv:2009.13295*.

Suniah Ayub, Salvatore T Scali, Julie Richter, Thomas S Huber, Adam W Beck, Javairiah Fatima, Scott A Berceli, Gilbert R Upchurch, Dean Arnaoutakis, Martin R Back, et al. 2019. Financial implications of coding inaccuracies in patients undergoing elective endovascular abdominal aortic aneurysm repair. *Journal of Vascular Surgery*, 69(1):210–218.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

JM Dixon, CP Armstrong, SW Duffy, RA Elton, and GC Davies. 1984. Upper gastrointestinal bleeding. a significant complication after surgery for relief of obstructive jaundice. *Annals of surgery*, 199(3):271.

Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Daniel Gartner, Rainer Kolisch, Daniel B Neill, and Rema Padman. 2015. Machine learning approaches for early drg classification and resource allocation. *INFORMS Journal on Computing*, 27(4):718–734.

Yong He, Cheng Wang, Nan Li, and Zhenyu Zeng. 2020. Attention and memory-augmented networks for dual-view sequential learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 125–134.

Yong He, Cheng Wang, Shun Zhang, Nan Li, Zhaorong Li, and Zhenyu Zeng. 2022. Kg-mtt-bert: Knowledge graph enhanced bert for multi-type medical text classification. *arXiv preprint arXiv:2210.03970*.

Md Mohaimenul Islam, Guo-Hung Li, Tahmina Nasrin Poly, and Yu-Chuan Li. 2021. Deepdrg: Performance of artificial intelligence model for real-time prediction of diagnosis-related groups. In *Healthcare*, volume 9, page 1632. MDPI.

Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. 2018. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39.

Mutahira Khalid, Hasan Ali Khattak, Arsalan Ahmad, and Syed Ahmad Chan Bukhari. 2022. Explainable prediction of medical codes through automated knowledge graph curation framework. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 331–336. IEEE.

Bong Sik Matthew Kim, Bob T Li, Alexander Engel, Jaswinder S Samra, Stephen Clarke, Ian D Norton, and Angela E Li. 2014. Diagnosis of gastrointestinal bleeding: A practical guide for clinicians. *World journal of gastrointestinal pathophysiology*, 5(4):467.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2021. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ digital medicine*, 4(1):103.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022. Hierarchical label-wise attention transformer model for explainable icd coding. *Journal of biomedical informatics*, 133:104161.

Guillermo López-García, José M Jerez, Nuria Ribelles, Emilio Alba, and Francisco J Veredas. 2023. Explainable clinical coding with in-domain adapted transformers. *Journal of Biomedical Informatics*, 139:104323.

Taulant Muka, David Imo, Loes Jaspers, Veronica Colpani, Layal Chaker, Sven J van der Lee, Shanthi Mendis, Rajiv Chowdhury, Wichor M Bramer, Abby Falla, et al. 2015. The global impact of non-communicable diseases on healthcare spending and national income: a systematic review. *European journal of epidemiology*, 30:251–277.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen.

2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(suppl_2):W170–W173.

DK Ousterhout and Irving Feller. 1968. Occult gastrointestinal hemorrhage in burned patients. *Archives of Surgery*, 96(3):420–422.

Kamal raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. 2022. Biosimcse: Biomedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4596–4608.

Owen Trigueros, Alberto Blanco, Nuria Lebeña, Arantza Casillas, and Alicia Pérez. 2022. Explainable icd multi-label classification of ehrs in spanish with convolutional attention. *International Journal of Medical Informatics*, 157:104615.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. 2022. A novel framework based on medical concept driven attention for explainable medical code prediction via external knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1407–1416.

SA Zafirah, Amrizal Muhammad Nur, Sharifa Ezat Wan Puteh, and Syed Mohamed Aljunid. 2018. Potential loss of revenue due to errors in clinical coding during the implementation of the malaysia diagnosis related group (my-drg®) casemix system in a teaching hospital in malaysia. *BMC health services research*, 18(1):1–11.