Systematic Comparative Analysis of Large Pretrained Language Models on Contextualized Medication Event Extraction

Tariq Abdul-Quddoos, Xishuang Dong, and Lijun Qian, *Senior Member, IEEE*

Abstract—Attention-based models have become the leading approach in modeling medical language for Natural Language Processing (NLP) in clinical notes. These models outperform traditional techniques by effectively capturing contextual representations of language.

In this research a comparative analysis is done amongst pre-trained attention based models namely Bert Base, BioBert, two variations of Bio+Clinical Bert, RoBerta, and Clinical Longformer on task related to Electronic Health Record (EHR) information extraction. The tasks from Track 1 of Harvard Medical School's 2022 National Clinical NLP Challenges (n2c2) are considered for this comparison, with the Contextualized Medication Event Dataset (CMED) given for these task. CMED is a dataset of unstructured EHR's and annotated notes that contain task relevant information about the EHR's. The goal of the challenge is to develop effective solutions for extracting contextual information related to patient medication events from EHR's using data driven methods.

Each pre-trained model is fine-tuned and applied on CMED to perform medication extraction, medical event detection, and multi-dimensional medication event context classification. Processing methods are also detailed for breaking down EHR's for compatibility with the applied models. Performance analysis has been carried out using a script based on constructing medical terms from the evaluation portion of CMED with metrics including recall, precision, and F1-Score. The results demonstrate that models pre-trained on clinical data are more effective in detecting medication and medication events, but Bert Base, pre-trained on general domain data showed to be the most effective for classifying the context of events related to medications.

Index Terms—Natural Language Processing, Electronic Health Records, Transformer, Medication, Bidirectional Encoder Representations from Transformers (BERT)

I. INTRODUCTION

An Electronic Health Record (EHR) holds records of patient health information generated by encounters in any care delivery setting, including information on patient demographics, progress notes, problem lists, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports [1]. The application of machine learning on EHR's serves to place the burden of extracting information on data driven models rather than people. The National Clinical NLP Challenges (N2C2), hosted by Harvard Medical School, focuses on the study of applying data driven approaches to mining clinical information holding a number of challenges since 2006. Track 1 of the 2022 N2C2 has task researchers with developing solutions for extracting information related to the context of medication mentions in EHR's. For this challenge the Contextualized Medication Event Dataset(CMED) has been released and is a dataset capturing relevant context needed to understand medication changes in clinical narratives [2], containing EHR's and annotated notes with task relevant information. In the competition, a number of methods were used with large language transformer based architectures, with some popular choices being Bert, RoBerta, DeBerta, Longformer, and Text-To-Text Transfer Transformer. In combination with these models additional architecture add-on methods consisted of multiple linear layers with combined losses and knowledge graphs [3].

In this work several pre-trained attention based language models undergo comparative analysis with information extraction task on CMED. The specific applied models are Bert Base, BioBert [4], two variations of Bio+Clinical Bert [5], RoBerta Base [6], and Clinical LongFormer [7] with pre-training from general, biomedical, and clinical domain corpora. The models are fine-tuned on CMED and used for the three task listed in section II. This study examines the performance of the applied models on the CMED dataset along with the differences in performance with regard to architecture differences, pre-training methods, and pre-training corpora. Pre- and post-processing methods are also of presented in this study given the wide range of syntactic structures used in EHR's, it can be difficult to judge how best the process an entire record. The contributions of this work are as follow: 1) A comparative analysis of pre-trained large langauge attention models performance of EHR related task; 2) Fine-tuning pre-trained models on CMED training data for medication detection,medication event detection, and multi-dimensional medication event context classification; 3) Processing methods for breaking down EHR's into format compatible with applied models.

The rest of this paper is organized as follows. In Section II the three modeling task are detailed, in section III the applied modeling achitecture's are detailed along with the differnces between applied pre-trained models, in section IV an overview of processing methods is given and results for experiments are analyzed, in section V an overview of related works is given, and in section VI this work is summarized and future work of interest is detailed.

X. Dong, T. Abdul-Quddoos, and L. Qian are with the Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT Center), Department of Electrical and Computer Engineering, Prairie View A&M University, Texas A&M University System, Prairie View, TX 77446, USA. Email: xidong@pvamu.edu, tabdulquddoos@pvamu.edu, liqian@pvamu.edu

II. 2022 National Clinical NLP Challenges

Track 1 of the 2022 N2C2 has task researchers with developing solutions for generating information related to the context of medication mentions in EHR's using data driven methods. A description of each task is given below.

A. Task 1: Medication Detection

Task 1 for the 2022 N2C2 has task researchers with identifying medication mentions in EHR's. This task has been well studied by NLP researchers with it also being a task in the 2018 N2C2 [8]. Knowing what medications and where they are in EHR's are an essential step for further medication related information extraction.

B. Task 2: Medication Event Classification

Task 2 for the 2022 N2C2 task researchers with classifying identified medication mentions from task 1 as having events associated with them. An event refers to any change that has to do with a particular medication within its context. The three classes for this task are given as disposition, no disposition, and undetermined. Disposition refers to an event occurring with the associated medication, no disposition refers to no event occurring for the associated medication, and undetermined refers to if annotators cannot determine if an event has occurred or not [2].

C. Task 3: Multi-dimensional Medication Event Context Classification

Task 3 for the 2022 N2C2 task researchers with classifying the context of medications that have been labeled with disposition in task 2. The context is classified across 5 dimensions, those dimensions being action, temporality, certainty, and actor, and negation [2]. Action refers to type of change being discussed and has the following 7 classes: Start, Stop, Increase, Decrease, Unique Dose, Other Change, and Unknown. Temporality refers to when the change occurred and has the following 4 classes: Past, Present, Future, Unknown. Certainty refers to if a change was implemented or just discussed and has the following 4 classes: Certain, Hypothetical, Conditional, and Unknown. Actor refers to who initiated the change and has the following and has the following 3 classes: Physician, Patient, Unknown. Negation refers to if the medication event is negated and has the following two classes: negated and not negated.

III. Method

In this work pre-trained attention based large language encoders are the primary modeling component studied. Each pre-trained encoder can be easily fine-tuned by switching the output layer once weights are initialized in pre-training to one suitable for the applied task. All models are pre-trained on a Masked Language Modeling(MLM) task allowing them to capture general representations of language distributions, significantly reducing training time for researchers and increasing performance on downstream task. The modeling pipeline is shown in figure 1.Words are first converted to subword tokens with each pre-trained model having its own specific subword

tokenizer. The two specific subword tokenization schemes applied between encoders are wordpiece tokenization and byte-pair encoding tokenization. Subword tokens are converted to embeddings and summed together with positional embeddings and encoded. For task 1 & 2 The encoded embeddings are placed into a feed-forward network and soft-max layer for label probabilities. For task 3 support vector machine are used to get label predictions instead of a feed-forward network with a soft-max.
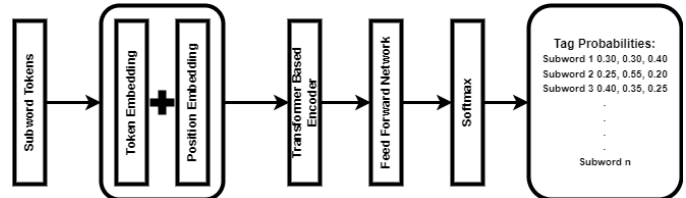

Fig. 1. Modeling Pipeline

All encoder implementations are imported from the the Transformers library [9] with PyTorch as the modeling framework. Specifics about pre-training and differences for each applied encoder are shown below.

1) Bert Base: The Bert base model applied is pre-trained on general domain corpora with all words made lowercase with fine-tuning on CMED done the same way. Bert Base contains a total of 110M parameters, made up of 12 transformer blocks each with 12 self-attention heads and hidden size for all layers in the model is 768 [10]. The tokenization scheme applied for all Bert Base varaints(BioBert & Bio+Clincal Bert's) follow wordpiece tokenization. Along with MLM, a next sentence prediction task is also applied as a pre-training task. The tokenization scheme applied for this encoder is wordpiece tokenization. The model was pre-trained on the following general domain corpora:

- BooksCorpus - 800M Words
- English Wikipedia - 2.5M Words

2) BioBert: BioBert is the first domain-specific Bert based model pre-trained on biomedical data [4]. The applied model is pre-trained with all lower case words so fine-tuning on CMED is done the same way. The tokenization scheme applied for this encoder is wordpiece tokenization. BioBert uses the previously explained Bert base model weights and does additionally pre-training on the following biomedical corpora:

- PubMed Abstracts - 4.5B Words
- PMC Full-text articles - 13.5B Words

3) Bio+Clinical Bert: Bio+Clinical Bert [5] is another domain specific Bert model pre-trained on clinical notes. The applied model is pre-trained with all lower case words so fine-tuning on CMED is done the same way. The tokenization scheme applied for this encoder is also wordpiece tokenization. Bio+Clinical Bert uses the previously explained BioBert model weights and undergoes further pre-training on the following clinical corpus.

- MIMIC-III v1.4 database - 2M Clinical Notes

There are two variation of Bio+Clinical Bert, one trained on all MIMIC-III notes and another only on MIMIC-III discharge

summaries, these will be referred to as Bio+Clinical Bert All Notes and Bio+Clinical Bert Discharge.

4) RoBerta Base: Robust Optimized Bert Approach known as RoBerta [6] holds the same architecture of the Bert Base but undergoes modified pre-training. The modifications are as follows, training the model longer with bigger batches over more data, removing the next sentence prediction task, training on longer sequences, and dynamically changing the masking pattern in MLM [6]. The tokenization scheme also differs from Bert following Byte-Pair encoding. The casesing of all words is kept in pre-training, with fine-tuning on CMED done the same way. RoBerta is pre-trained with the following corpora:

- BooksCorpus - 800M Words
- English Wikipedia - 2.5M Words
- CC-News - 60M News Articles
- OpenWebText - 38GB of web content
- Stories - 2.1B Tokens

5) Clinical Longformer: Clinical Longformer [7] follows the Longformer Base [11] architecture with additional pre-training done on clinical text. Longformer is a Bert like architecture that allows for input sequences of up to 4069 instead of 512 limit, although in this work 512 is still applied. Longformer base takes the Roberta base checkpoint and applies a modified attention mechanism with a local windowed attention along with the global attention [11]. Like RoBerta the tokenization scheme applied follows Byte-Pair encoding. Pre-training is done with keeping the case of all words, so fine-tuning on CMED is done the same way. Along with the corpora from Roberta, Clinical Longformer is pre-trained with the following corpora:

- Realnews Corpora - 1.8B Tokens
- MIMIC-III v1.4 database - 2M Clinical Notes

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Contextualized Medication Event Dataset

The Contextualized Medication Event Dataset(CMED) has been released for track 1 of the 2022 n2c2 and is a dataset that captures relevant context of medication changes documented in clinical notes [2]. CMED consist of 500 clinical notes & annotated notes with task relevant information, with 9,012 medication mentions across all the notes. The clinical notes from CMED are a portion of the data from the 2014 i2b2/UTHealth Natural Language Processing shared task [12]. The class distributions of CMED for all three task are shown in tables I, II, & III.

TABLE I
CMED TASK 1 DISTRIBUTION. ADAPTED FROM [3]

Label	Train	Test	Total
Drug	7229	1783	9012

TABLE II
CMED TASK 2 DISTRIBUTION. ADAPTED FROM [3]

Label	Train	Test	Total
Disposition	1412	335	1747
NoDisposition	5260	1326	6586
Undetermined	557	122	679

TABLE III
CMED TASK 3 DISTRIBUTION. ADAPTED FROM [3]

Context Dimension	Label	Train	Test	Total
Action	Start	568	131	699
	Stop	340	67	407
	Increase	129	22	151
	Decrease	54	13	67
	Unique Dose	285	88	373
	Other Change	1	0	1
	Unknown	35	14	49
Temporality	Past	744	173	917
	Present	494	132	626
	Future	145	29	174
	Unknown	29	1	30
Actor	Physician	1278	311	1589
	Patient	106	17	123
	Unknown	28	7	35
Certainty	Certain	1176	281	1457
	Hypothetical	134	33	167
	Conditional	100	15	115
	Unknown	2	6	8
Negation	Negated	32	6	38
	Not Negated	1380	329	1709

B. Processing

The pre-processing pipeline is shown in figure 2 and in the following order consist of annotation parsing, EHR tagging, EHR sentence/section segmentation, word tokenization, and subword tokenization. All pre-processing is done with Python and is largely rule-based with the exception of sentence segmentation and word tokenization being done using pre-trained models from the PunkT [13] package and wordpiece tokenization or byte-pair tokenization done with pre-trained models paired with each applied encoder. For task 1 and task 2 the data is pre-processed into a BIO tag format for Named Entity Recognition. For task 3 a BIO format is not used, since only medications that have a disposition label from task 2 are used for training data, these medications are given just their class labels for each context dimension.
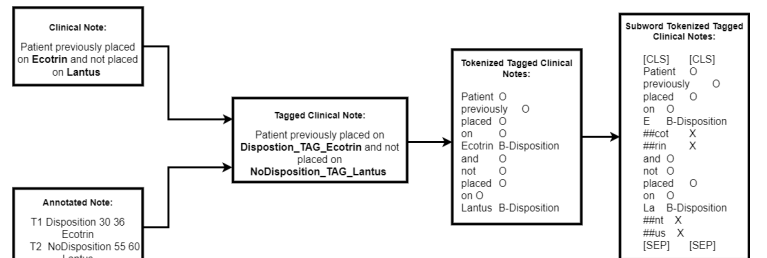

Fig. 2. Pre-Processing Pipeline

The post-processing pipeline is shown in figure 3 and consist of word reconstruction from subwords and placing predictions in an annotated format for compatibility with an evaluation script released for the 2022 N2C2 competition.
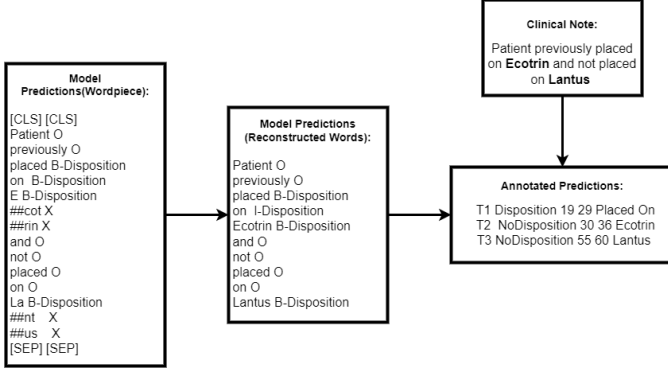

Fig. 3. Post Processing Pipeline

For training in task 1 and task 2 the entire model with encoder and output layers are fine-tuned with CMED with all words from the training portion used. For task 3 only the SVM's are trained using the embeddings from each pre-trained encoder, with an SVM for each context dimension.

C. Evaluation Metrics

The evaluation metrics used in this study are precision, recall and F1-Score. Precision is defined as the probability that an object is relevant given that it is returned by the system [14] or as the number of true positives(correct labels) for a particular label out of all entities returned by a system and is shown in equation 1.

$$Precision = \frac{TP}{TP + FP}.$$ (1)

Recall is defined as the probability that a relevant object is returned by a system [14] or the number of true positives(correct labels) for a particular label out of all the entities that actually have that been returned by a system and is shown in equation 2.

$$Recall = \frac{TP}{TP + FN}.$$ (2)

An F-Score is a measure of a models accuracy on a dataset and is defined as the harmonic mean between precision and recall and is shown in equation 3,

$$FScore = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$ (3)

Predictions are based on character positions in EHR's and strict and lenient matching are considered for calculating each evaluation metrics, where the strict matching refers to if an prediction matches the exact correct character position associated with the medication, while with lenient matching if there is overlap in the prediction and correct character positions, then the prediction counts as correct.

D. Results and Analysis

1) Task 1: Medication Detection: Results for task 1 are shown in table IV. On average models achieve a strict F-Score of 0.9226 and lenient F-Score of 0.9581. Bio+Clinical Bert Discharge shows the best performance for this task with a strict score 0.0129 higher than the average at 0.9355 and lenient score .0089 higher than the average at 0.9669. RoBerta base shows the worst performance for this task with a a strict F-Score of 0.9075, which is .0151 less than the average and lenient F-Score of 0.9476 which is 0.0105 less than the average. Across all metrics models pre-trained on clinical data(Bio+Clinical Bert's and Clinical Longformer) show the best performance, while models trained on only general domain data achieve the worst performance(Bert Base and RoBerta Base).

2) Task 2: Medication Event Classification: Results for task 2 are shown in table V. On average models achieve a strict F-Score of 0.8280 and lenient F-Score of 0.8614. Clinical Longformer is the top performing model for this task with a strict F-Score of 0.8515, which is 0.0235 higher than the average and lenient F-Score 0.8793 which is 0.0179 higher than the average. Like task 1, the model pre-trained with clinical data are the top performers across all metrics with specifically Clinical Longformer and Bio+Clinical Bert Discharge having the top scores for all metrics. Also like task 1 models pre-trained on general domain data only achieve the lowest performance on this task.

3) Task 3: Multi-dimensional Context Classification: Results for task 3 are shown in table VI, Strict and lenient scores are not considered for this task as position of medications are not being predicted for this task. Results are also taken from two perspectives, the first like the other task is the overall F-score across context dimensions, the second perspective considers all context dimensions predictions combined and counts prediction to be correct if all dimensions have correct labels. For the overall performance on average models achieve and F-Score of 0.5986 and for the combined performance models achieve an average score of 0.1754. Bert Base shows the best performance across all categories largely outperforming other models with an overall score of 0.7387, which is 0.1401 higher than the average and a combined F-Score of 0.3006, which is 0.1251 higher than the average. RoBerta and Clinical Longformer show the poorest performance, largely under-performing all Bert Base variants.

V. RELATED WORKS

Since the release of the transformer [15] model in 2017 a number of studies have been done around their application on clinical text. The GPT variant GPT2 has been applied for generating synthetic EHR data as a solution to overcoming privacy concerns with EHR's [16]. Bert has been popular in information extraction, where like in this work the encoder is fine-tuned on an EHR dataset. This strategy has been applied in task such as predicting depression [17], identifying events where patients have injuries due to falls [18], and identifying adverse drug events [19]. Along with fine-tuning transformer models for task, several models like Clinical Longformer and

TABLE IV
PERFORMANCE COMPARISON ON MEDICATION DETECTION

Model	Strict Evaluation Performance			Lenient Evaluation Performance		
Pretrained Model	Precision	Recall	F-Score	Precison	Recall	F-Score
$BERTbase$	0.8939	0.8365	0.9147	0.9297	0.9739	0.9513
$BioBERT$	0.9154	0.9235	0.9239	0.9521	0.9700	0.9610
$Bio+ClinicalBertAllNotes$	**0.9281**	0.9291	0.9255	**0.9612**	0.9688	0.9650
$Bio+ClinicalBertDischarge$	0.9252	0.9461	**0.9355**	0.9562	**0.9779**	**0.9669**
$RoBerta$	0.8925	0.9229	0.9075	0.9320	0.9637	0.9476
$ClinicalLongformer$	0.9097	**0.9478**	0.9284	0.9374	0.9768	0.9567

TABLE V
PERFORMANCE COMPARISON ON MEDICATION EVENT CLASSIFICATION

Model	Strict Evaluation Performance			Lenient Evaluation Performance		
Pretrained Model	Precision	Recall	F-Score	Precision	Recall	F-Score
$BERTbase$	0.7922	0.8295	0.8104	0.8268	0.8657	0.8458
$BioBERT$	0.8125	0.8272	0.8198	0.8464	0.8618	0.8540
$Bio+ClinicalBertAllNotes$	0.8268	0.8329	0.8298	0.8650	0.8714	0.8682
$Bio+ClinicalBertDischarge$	**0.8370**	0.8555	0.8462	**0.8659**	0.8850	0.8753
$RoBerta$	0.7971	0.8283	0.8103	0.8322	0.8601	0.8459
$ClinicalLongformer$	0.8346	**0.8691**	**0.8515**	0.8618	**0.8975**	**0.8793**

TABLE VI
PERFORMANCE COMPARISON ON MULTI-DIMENSIONAL CONTEXT CLASSIFICATION

Model	Overall Performance			Combined Performance		
Pretrained Model	Precision	Recall	F-Score	Precision	Recall	F-Score
$Bertbase$	**0.7569**	**0.7188**	**0.7387**	**0.3091**	**0.2925**	**0.3006**
$BioBert$	0.7104	0.6722	0.6908	0.2366	0.2239	0.2301
$Bio+ClinicalBertAllNotes$	0.6826	0.6460	0.6638	0.2303	0.2179	0.2239
$Bio+ClinicalBertDischarge$	0.6650	0.6293	0.6466	0.2114	0.2000	0.2055
$RoBerta$	0.4038	0.3821	0.3926	0.0410	0.0388	0.0399
$ClinicalLongformer$	0.4719	0.4466	0.4589	0.0536	0.0507	0.0521

Bio+Clinical Bert have been pre-trained with masked langauge modeling to create domain specific models. Some of these pre-trained models include MedBert [20], CancerBert [21], & BlueBert [22].

Before transformer based models Recuurent Neural Network models were the standard approach for modeling EHR's. Track 2 of the 2018 N2C2 like track 1 of the 2022 N2C2 involved medication related information extraction and the RNN variant Bidirectional long short-term memory with Condition Random Fields(BiLSTM CRF's), was a popular modeling choice, with 9 of the top teams incorporating them in their system. [23]. BiLSTM-CRFs use a BiLSTM to create a series of state representations that are then used as input into a CRF for labeling [23]. Static embedding methods had been applied to accompany the BiLSTM-CRF's ,one popular method was to use the MIMIC-III dataset to create pre-trained word embeddings with the Word2Vec package another was to use GloVe embeddings[24].

VI. CONCLUSION

In this work a comparative analysis is presented amongst attention based large language models for the following EHR information extraction task: medication detection, medication event classicatition, and multi-dimensional medication event context classification. The applied models are Bert Base, BioBert, 2 variations of Bio+Clinical Bert, RoBerta Base, and Clinical LongFormer. The models are evaluated on all task with recall, precision, and F-Score as metrics. For medication

detection models pre-trained on clincial data show the best performance with Bio+Clinical Bert pre-trained on MIMIC III Discharge notes acheiveing the highest performance with a strict F-Score of 0.9355 and lenient F-Score of 0.9669. For medication event classication, like medication detection, models pre-trained with clinical data achieve the best performance with the highest F-Scores coming from Clinical LongFormer with a strict F-Score of 0.8515 and lenient F-Score of 0.8793. For multi-dimensional medication event context classification, Bert base achieves the best performance. For this task two perspectives are considered, the first is the overall F-Score, for this Bert Base achieves a score of 0.7387, the second perspective is scoring on how many predictions have all context dimension prediction correct for a medication, and for this Bert Base achieves a score of 0.3091. For future work, increasing performance on the third task is considered, especially for generating correct predictions across all dimensions. Methods around data augmentation are of interest, there is a lack of data for this task, with some classes having as little as 1 label in the training set.

contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) or the U.S. Government.

REFERENCES

[1] V. Ehrenstein, H. Kharrazi, and H. Lehmann et al., *Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition*. Rockville: Agency for Healthcare Research and Quality, 2019.

[2] D. Mahajan, J. Liang, and C. Tsou, "Toward understanding clinical context of medication change events in clinical narratives," pp. 1–8, 2022.

[3] D. Mahajan, J. Liang, C. Tsou, and O. Uzuner, "n2c2 2022 challenge: Track 1 contextualized medication event extraction," "Presented as the 2023 N2C2 Workshop, Washington DC", 2023.

[4] J. Lee, W. Yoon, and S. Kim et al., "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[5] E. Alsentzer, J. Murphy, and W. Boag et al., "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, (Minneapolis, Minnesota, USA), pp. 72–78, Association for Computational Linguistics, June 2019.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[7] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, "Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences," *arXiv preprint arXiv:2201.11838*, 2022.

[8] A. Stubbs, M. Filannino, and E.Soysal et al., "Cohort selection for clinical trials: n2c2 2018 shared task track 1," *Journal of the American Medical Informatics Association*, vol. 26, pp. 1163–1171, 09 2019.

[9] T. Wolf, L. Debut, and V.Sanh et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[10] J.Devlin, M.Chang, K.Lee, , and K.Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *Clinical Orthopaedics and Related Research*, vol. 1810.04805, 2018.

[11] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv:2004.05150*, 2020.

[12] V. Kumar, A. Stubbs, S. Shaw, and O. Uzuner, "Creation of a new longitudinal corpus of clinical narratives," *Journal of Biomedical Informatics*, vol. 58, no. 5, pp. S6–S10, 2015.

[13] T.Kiss and J.Strunk, "Unsupervised Multilingual Sentence Boundary Detection," *Computational Linguistics*, vol. 32, pp. 485–525, 12 2006.

[14] C. Goutte and E. Gaussier, "A probailistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval*, pp. 345–359, 2005.

[15] A. Vaswani, N. Shazeer, and N. Parmar et al., "Attention is all you need," in *31st Conference on Neural Information Processing*, 2017.

[16] C. Libbi, J. Trienes, D. Trieschnigg, and C. Seifert, "Generating synthetic training data for supervised de-identification of electronic health records," *Future Internet*, vol. 13, no. 136, 2021.

[17] Y. Meng, W. Speier, M. K. Ong, and C. W. Arnold, "Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3121–3129, 2021.

[18] S. Fu, B. Thorsteinsdottir, X. Zhang, G. S. Lopes, S. R. Pagali, N. K. LeBrasseur, A. Wen, H. Liu, W. A. Rocca, J. E. Olson, J. S. Sauver, and S. Sohn, "A hybrid model to identify fall occurrence from electronic health records," *International Journal of Medical Informatics*, vol. 162, p. 104736, 2022.

[19] X. Yang, Z. Yu, Y. Guo, J. Bian, and Y. Wu, "Clinical relation extraction using transformer-based models," *CoRR*, vol. abs/2107.08957, 2021.

[20] C. Vasantharajan, K. Tun, and H. Thi-Nga et al., "Medbert: A pre-trained language model for biomedical named entity recognition," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1482–1488, 2022.

[21] S. Zhou, N. Wang, L. Wang, H. Liu, and R. Zhang, "CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records," *Journal of the American Medical Informatics Association*, vol. 29, pp. 1208–1216, 03 2022.

[22] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," in *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pp. 58–65, 2019.

[23] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 3–12, 2020.

[24] H. Dai, C. Su, and C. Wu, "Adverse drug event and medication extraction in electronic health records via a cascading architecture with different labeling models and word embeddings," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 47–55, 2020.