applied sciences

MDPI

Article

ConBERT: A Concatenation of Bidirectional Transformers for Standardization of Operative Reports from Electronic Medical Records

Sangjee Park [1,†], Jun-Woo Bong [2,†], Inseo Park [1], Hwamin Lee [3], Jiyoun Choi [4], Pyoungjae Park [2], Yoon Kim [5,6,7], Hyun-Soo Choi [5,6,7,*] and Sanghee Kang [2,*]

1 Department of Convergence Security, Kangwon National University, Chuncheon 24253, Korea
2 Department of Surgery, Korea University Guro Hospital, Korea University College of Medicine, Seoul 08308, Korea
3 Department of Medical Informatics, Korea University College of Medicine, Seoul 02841, Korea
4 Korea University Guro Hospital Medical Record Information Team, Seoul 08308, Korea
5 Department of Computer Science, Engineering, Kangwon National University, Chuncheon 24253, Korea
6 Department of Medical Bigdata Convergence, Kangwon National University, Chuncheon 24253, Korea
7 Ziovision, Chuncheon 24341, Korea
* Correspondence: choi.hyunsoo@kangwon.ac.kr (H.-S.C.); kasaha1@korea.ac.kr (S.K.)
† These authors contributed equally to this work.

Abstract: This operative report documents the details of a surgery. Standardization of the medical terminology for the operative report written in free text is significant for performing medical research and establishing insurance systems by accurately sharing information on treatment. However, standardization of operative reports is a labor-intensive task that has a risk of induced errors. We have proposed a concatenation of bidirectional encoder representations from transformers (ConBERT) model for predicting the International Classification of Disease-9 code using the operative report and diagnosis recorded in free text to standardize the operative report automatically. We compared the pre-trained models of BERT and character BERT and created a new model by concatenating the combinations of each model. The proposed ConBERT model showed a micro AP score of 0.7672, F1 score of 0.7415, and AUC of 0.9842. In addition, we developed a web-based application to demonstrate the performance of our model and make it publicly accessible.

Keywords: standardization of operative report; medical code prediction; ConBERT; ICD-9 code

check for updates

Citation: Park, S.; Bong, J.-W.; Park, I.; Lee, H.; Choi, J.; Park, P.; Kim, Y.; Choi, H.-S.; Kang, S. ConBERT: A Concatenation of Bidirectional Transformers for Standardization of Operative Reports from Electronic Medical Records. *Appl. Sci.* **2022**, *12*, 11250. https://doi.org/10.3390/app122111250

Academic Editor: Enno van der Velde

Received: 12 September 2022
Accepted: 2 November 2022
Published: 6 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remarkable and steady developments in medical research and tools have led to continuous improvements in surgical treatment. Currently, various surgeries are performed in different fields. In general surgery, for example, an identical procedure can be performed using either an open approach or minimally invasive tools such as laparoscopy and robotic systems [1]. Additionally, different types of operative reports for rectal cancer, including low anterior resection or ultra-low anterior resection, can be performed for different individuals, depending on the distance from the anus to the location of rectal cancer [2]. Therefore, exact documentation of operation records is essential for providing appropriate information.

The operative report, which denotes a representative procedure such as "appendectomy" or "cholecystectomy" is an important component of the surgical record because it contains key surgical information. The operative report is dictated immediately after an operation and is included in the patient's electronic health record. It contains important information, such as the type of surgery and the method of approach. Surgical records are primarily created by surgeons; however, their documentation varies substantially. Surgeons generally record operative reports in free text and use non-standard abbreviations (for example, appendectomy, laparoscopic appendectomy, lap. appendectomy, lapa.

appe, appendectomy, and laparoscopic). Nonstandard terminology in operative reports can significantly affect data quality and confuse physicians. Therefore, the standardization of operative reports is an important topic to be analyzed as it plays a critical role in clinical research and quality assurance in healthcare. The International Classification of Diseases (ICD), which is a database comprising clinical terminologies managed by the World Health Organization, is widely accepted as a standardized medical terminology guide [3]. Standardization of operative reports implies matching unstructured data from operative reports written in free text with structured data such as ICD codes. However, the standardization of operative reports requires manual chart review, which is a time- and effort-consuming process.

Recently, active research [4–8] is being conducted to automate ICD-9 codes. For this purpose, most studies use clinical notes and discharge summaries, which are information-intensive data sources. To the best of our knowledge, no study has attempted to standardize ICD-9 codes using data sources containing lesser information, such as operative reports and diagnoses. Hence, we considered a bidirectional encoder representation from transformers (BERT)-based model to predict the ICD-9 code using operative reports and diagnoses. BERT [9] demonstrated remarkable performance in various natural language processing fields. However, BERT trained on general domain corpora showed relatively low performance on medical data. Therefore, BERTs trained in biomedical corpora are provided separately. In addition, given the nature of clinical data, there are frequent typographical errors, new words, and abbreviations, and the current word-level BERT is unsuitable. Therefore, character BERT, also known as character-level BERT, has been proposed [10]. Although it efficiently managed a small amount of data, there were errors/limitations when handling a large amount of data. Therefore, we propose a concatenation of BERT (ConBERT) model that utilizes both word- and character-level BERT for ICD-9 code automation. The proposed model showed AP score of 0.7672, F1 score of 0.7415, and AUC of 0.9842 for ICD-9 code automation. We have publicly provided this service as a web-based application accessible at http://opti.ziovision.ai/ (accessed on 8 September 2022).

2. Materials and Methods

2.1. Clinical Data

From the electronic medical records of two tertiary referral hospitals, the Korea University Anam Hospital and the Korea University Guro Hospital, which are independent organizations, we collected all the surgical records of patients who underwent surgery in the Department of Surgery between January 2009 and December 2020. The surgical records included patient ID, date, division, surgeon details, preoperative diagnosis, postoperative diagnosis, operative report, operative findings, and other information (including blood loss, complications during surgery, and anesthesia). Data on the date, division, postoperative diagnosis, and operative reports were extracted from all the surgical records to establish a dataset (Table 1). The division included data on the type of surgery among the eight sub-department components of the Department of Surgery: breast and endocrine surgery, gastroesophageal surgery, hepatobiliary surgery, colorectal surgery, pediatric surgery, transplantation and vascular surgery, general surgery, and acute care surgery. The postoperative diagnosis was recorded in free text. The operative report comprised three records of the original text and an ICD-9-matched code and name. The Medical Record Information Team in the Korea University Guro Hospital, performed all the works of labelling and matching the original text information with standardized data using their own dictionaries. Distinct datasets from the two institutions were merged into one. The sample sizes of the two independent datasets were 45,211 cases from the Korea University Anam Hospital and 35,862 cases from the Korea University Guro Hospital.

Table 1. Surgical records in the dataset.

ID	Date	Division	Postoperative Diagnosis	Operative Report (Original)	Operative Report (ICD-9)	Code (ICD-9)
1	6 August 2019	Hepatobiliary	GB stone	lap. cholecystectomy	Cholecystectomy; laparoscopic	51.23
2	21 May 2012	Colorectal	rectal ca. (AV 4 cm)	ULAR	Resection; rectum, other anterior	48.63
3	29 January 2020	Colorectal	r/o appendiceal cancer	Lap. RHC	Hemicolectomy; right Laparoscopy	45.73 54.21
4	22 June 2018	Endocrine and Breast	Rt. PTC	THYROIDECTOMY, TOTAL	Thyroidectomy; complete	06.4
				central LN dissection	dissection; neck, not otherwise specified, radical	40.40
5	19 August 2020	Transplantation and Vascular	HBV LC with HCC	LDLT	Transplant; liver, other	50.59

ICD = International Classification of Diseases.

2.2. Preprocessing

Four types of preprocessing were performed (Figure 1) to effectively analyze the data. First, for data cleansing, data with missing values and special characters, such as +, −, →, >, <, ≪, and ≫—that is, content-wise insignificant information—were removed. Additionally, specific numbering symbols, such as [1] and ①, used only in specific institutions, were removed. Second, we removed grammatically appropriate but contextually meaningless stop words using the Natural Language Toolkit [11], which is a text-processing library available in Python. Third, we created an input by combining the operative report and diagnosis. Finally, duplicate data from the same final input data and ICD-9 code were removed. After preprocessing, the number of data points obtained was 45,853. About 46% of the data in the dataset consisted of data with the exact same operative report and ICD-9 code. Therefore, if the data were used as it is, it could be biased to a specific value, so duplicate data were considered a single sample.

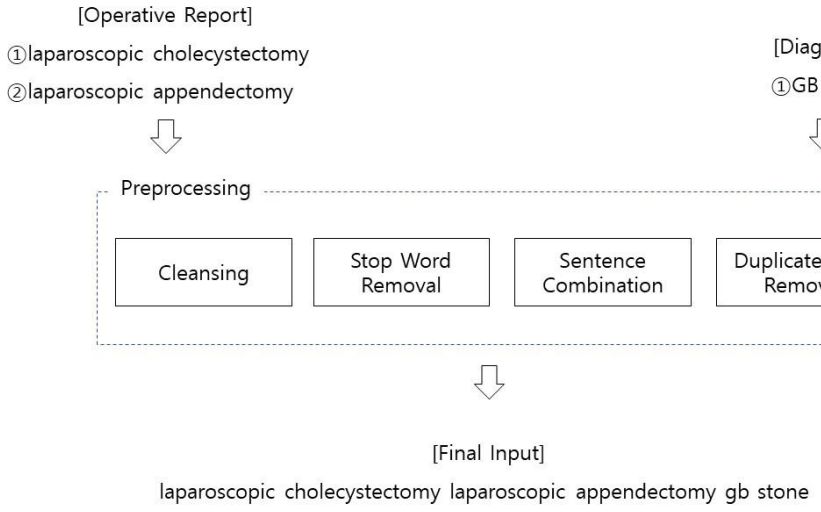


Figure 1. Preprocessing workflow.

In a unique case, which has only one data point with a specific label (ICD-9 codes such as 45.72, 48.63, and 54.21), the data cannot be divided into training and test sets, making it difficult to accurately measure performance. Therefore, we used data obtained after removing a small number of labels. In addition, multilabel data were used in the proposed model. Multilabel refers to the existence of one or multiple labels coexisting in a single dataset. In the D2SBERT [12] study, this problem was eliminated by using only the top 50 labels. In this study, we identified the removal rate that could be equally divided

into training and test sets through empirical experiments. Consequently, the data were removed for ICD-9 codes that appeared 11 or fewer times. The removed data, containing uncommon terms, accounted for 3% of the total data. After removal, 44,341 data points and 353 ICD-9 codes were obtained. The data were used in the multilabel classification library to ensure that the multilabel ICD-9 codes were balanced in the training and test sets. The multilabel classification library was used in scikit-learn [13], which is an open-source machine learning tool available in Python, and the training and test datasets were divided in an 8:2 ratio.

2.3. Methods

In this section, we discuss the ConBERT model, which is an ensemble model for achieving higher performance in ICD-9 code prediction.

2.3.1. Embedding

Embedding is a natural language processing technique that expresses words as numerical vectors. The generated vector contains semantic and grammatical information; therefore, various operations, such as calculating the similarity between words and sentences, are possible. In the case of ICD-9 code automation, there is a difference in the indication of the operative report and diagnosis, even with the same ICD-9 code. In this case, the similarity of expressions can be compared by embedding. Representative embedding models include Word2Vec [14], Glove [15], FastText [16], ELMo [17], and BERT. As BERT has demonstrated outstanding performance on various tasks compared with other models, we utilized BERT-based embedding. All the pre-trained models and the corresponding corpuses used for pre-training are listed in Table 2. A detailed description of each model is provided in the following sections.

Table 2. The pre-training model and learned corpus.

Model	Corpus
MedicalBERT [10]	MIMIC-III Clinical note, PMC OA biomedical paper abstract
UmlsBERT [18]	Intensive Care III (MIMIC-III), MedNLi, i2b2 2006, i2b2 2010, i2b2 2012, i2b2 2014
BioBERT [19]	English Wikipedia, BooksCorpus, PubMed Abstracts, PMC Full-text articles
MedicalCharacterBERT [10]	MIMIC-III Clinical note, PMC OA biomedical paper abstract

BERT = bidirectional encoder representation from transformers.

2.3.2. BERT

BERT, which was proposed by Google in 2018, outputs an embedding vector of a word according to a given context. Existing embedding models, such as Word2Vec, Glove, and Fast Text, ignore the bidirectionality of the context, which limits contextual understanding. In contrast, BERT considers bidirectionality and can better understand more complicated contexts. Additionally, BERT has achieved state-of-the-art performance in various downstream tasks through fine-tuning; however, it is difficult to estimate its performance on datasets containing biomedical texts. Moreover, the word distributions of general and biomedical corpora are quite different [19]. Consequently, we utilized pre-trained BERT for the biomedical corpora.

2.3.3. Character BERT

Whereas BERT only outputs a word-level embedding vector, Character BERT outputs a character-level embedding vector. BERT has a performance degradation problem on noisy datasets because it considers words with typographical errors as new words. In practice, operative reports and diagnoses have frequent typographical errors, various forms of abbreviations, and novel jargon. To address this problem, we used Character BERT, which uses character-level embedding, in addition to BERT. Character BERT exhibits robustness in

noisy datasets and novel words. The pre-trained models and corpuses used for pre-training are shown in Table 2.

2.3.4. Model Aggregation

We have proposed the ConBERT model, which is a concatenation BERT that combines word- and character-level BERT to improve performance. The vectors generated through the word- and character-level BERT were concatenated. The concatenated embedding vector was converted into a probability vector through a fully connected layer and a sigmoid layer. The final output was the ICD-9 code with a probability value above the threshold. We add a classification layer and a sigmoid layer to the end of the pre-trained BERT model to fine-tune the icd-9 code to classify. The classification layer used fully connected layer whose input size is the sum of output dimensions of each BERTs, and output dimension is the number of multi-label class, 353. The model architecture is illustrated in Figure 2.

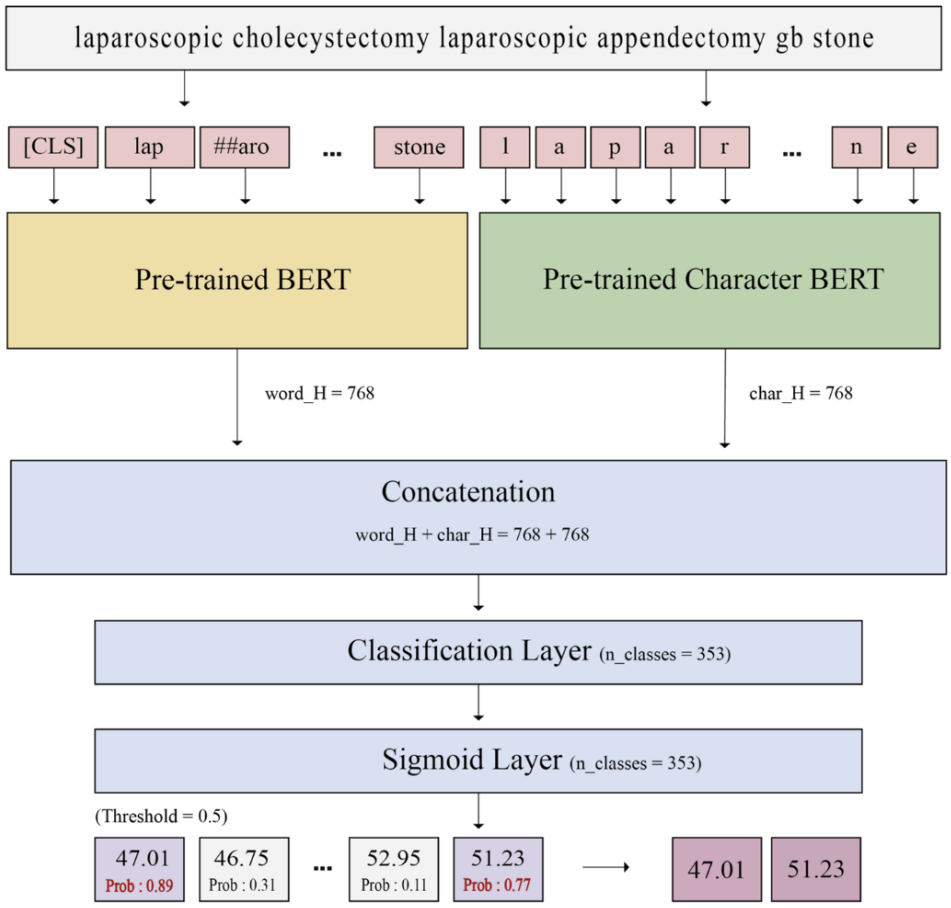


Figure 2. ConBERT model architecture.

2.3.5. Training Details

Training was performed using Pytorch Framework 1.11.0, Python 3.8.12, and NVIDIA RTX 3090Ti. The batch size was 128, optimizer was AdamW [20], and initial learning rate was 0.0001; the learning rate scheduler, warmup [21], and early stopping was used. The maximum sequence length of BERT was 256, and the maximum token length of Character BERT was 256. The entire architecture is trained using surgical data and binary cross-entropy loss for classification. In addition, we found hyperparameters such as learning rate through 10-fold validation. Found through learning rate finder of Trainer module provided by pytorch lightning [22]. The training and validation was conducted using 10-fold cross validation, and the training data and test data set at a ratio of 8:2 were used. In addition, trained 40 epochs for each fold and changed train dataset and validation dataset. After

that, the average of evaluation metrics was calculated for each model, and for final model all datasets except the test dataset were used as training data based on the model with the highest performance.

2.3.6. Evaluation

We evaluated the model with the F1 score and area under the curve (AUC) used in previous studies [23,24], and the average precision (AP), which compensates for the bias in precision and recall values according to threshold values. The AUC is the area below the receiver operating characteristic (ROC) curve. In the graph of the ROC curve, the X-axis is the false positive rate and the Y-axis is the true positive rate. The average precision (AP) score summarizes the precision-recall curve as a weighted average of precision at each threshold and uses the recall difference from the previous threshold as a weight.

3. Results

3.1. Input Comparison

Table 3 shows the data combining the operative report and diagnosis, and the comparison of each feature. Consequently, data combining the operative report and diagnosis showed improvements in all performance indicators. Training was confirmed by dividing the train dataset and the test dataset in a ratio of 8:2.

Table 3. Comparison of operative report and diagnosis data.

Proposed Model	AP		F1		AUC	
	Micro	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**
Diagnosis	0.4552	0.2347	0.4694	0.1736	0.9710	0.9282
Operative report	**0.7692**	0.4574	0.7292	0.4037	**0.9889**	0.9692
Operative report + Diagnosis	0.7647	**0.4891**	**0.7509**	**0.4382**	0.9812	**0.9675**

AUC, area under the curve.

3.2. Comparison of Pre-Trained Models

The prediction performances of the five pre-trained BERT models are listed in Table 4. The model with the highest performance is the Umls BERT model, with a very slight difference from the Bio BERT model, and shows an AP score of 0.7872, F1 score of 0.7603, and AUC of 0.9863 based on the micro-average setting. Unlike other models, base BERT is a trained model that uses a general corpus rather than a medical corpus. Therefore, as explained in Section 2.3.2, the performance of the tasks in this study was low. Additionally, the overall performance of the Medical Character BERT model was lower than that of the other word-level BERT models, as shown in Table 5. However, its performance on minority data was better than that of the other word-level BERT models.

Table 4. Comparison of models.

Model		AP		F1		AUC	
		Micro	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**
Single Model	* Base	0.7854	0.5718	0.7570	0.3210	0.9860	0.7565
	Umls	0.7872	0.5815	0.7603	0.3383	0.9863	0.7575
	Medical	0.7836	0.5786	0.7584	0.3333	0.9854	0.7564
	Bio	0.7895	0.5815	0.7592	0.3369	0.9863	0.7557
	* MC	0.7751	0.5445	0.7495	0.2965	0.9853	0.7566
Aggregated Model	Medical + * MC	0.7898	0.5891	0.7604	0.3578	0.9867	0.7580
	Umls + * MC	**0.7937**	**0.5959**	**0.7622**	**0.3643**	0.9876	**0.7587**
	Bio + * MC	0.7922	0.5905	0.7593	0.3637	**0.9877**	0.7583

* Base = base BERT: pre-trained with general corpuses, * MC = Medical Character BERT, AUC = area under the curve.

Table 5. Comparison of the F1 score by labels.

	ICD-9 Code	F1		Number of Data in Test Set
		Umls	Character	
Majority labels	51.23	**0.9324**	0.9245	628
	06.4	**0.9751**	0.9638	264
	50.12	**0.4096**	0.3871	111
Minority labels	46.43	0.7059	**0.7200**	27
	38.03	0.7600	**0.8077**	23
	45.52	0.7857	**0.8148**	12
	39.25	0.6000	**0.6667**	6

ICD, International Classification of Diseases.

3.3. Comparison of Aggregated Models

The results of the combinations of the three word-level BERT models (Umls, Medical, and Bio) and the Medical Character BERT model are shown in the model aggregation in Table 4. An aggregated model is better for macro performance compared with a single model without degradation of the micro performance. Therefore, we proposed a combination of Umls BERT and Medical Character BERT models, which exhibited the best macro F1 performance. This is because a high macro indicator indicates a good model for minority data.

The proposed model learned the entire data by dividing training and test dataset at a ratio of 8:2, and used a combination of the most performing Umls BERT and Medical Character BERT when learned using 10-fold. The performance measurement of the Aggregated Model was based on the highest training performance and the multi-label AUC calculation method, but the final performance was based on the average performance and the overall AUC for comparison with previous papers. The final performance of the proposed model is shown in Table 6 with a micro-based AP score of 0.7672, an F1 score of 0.7415 and an AUC of 0.9842.

Table 6. Test performance of Proposed Models.

Model		AP		F1		AUC	
		Micro	Macro	Micro	Macro	Micro	Macro
Proposed Model	Umls + * MC	0.7672	0.4899	0.7415	0.3975	0.9842	0.9703

* MC = Medical Character BERT.

A comparison of Umls BERT with Character BERT based on the micro F1 score reveals that the performance of Umls BERT is better than that of Character BERT on majority data, and the performance of Character BERT is better on minority data.

3.4. Web-Based Application

Based on this model, we developed a web application. The front-end was implemented using Vue.js, and the model server was implemented using TorchServe. The model used for the web application was trained using all the data. The classification speed of the model processes 34 data per second, and the web application is capable of both single prediction and multi-prediction. On the single prediction page (Figure 3), the corresponding ICD-9 code and probability can be obtained by entering the name of the surgery and diagnosis. On the multi-prediction page (Figure 4), the ICD-9 codes and probabilities for multiple operative reports and diagnoses that can be obtained through the CSV file upload are presented.

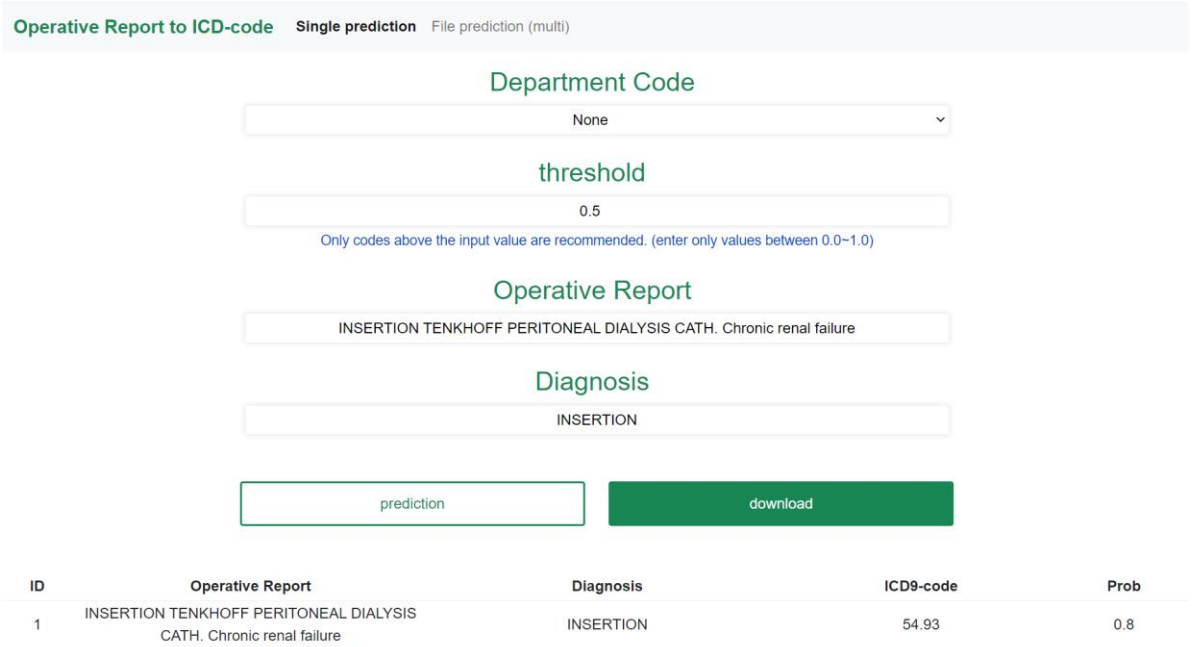


Figure 3. Single prediction page of the web-based application.

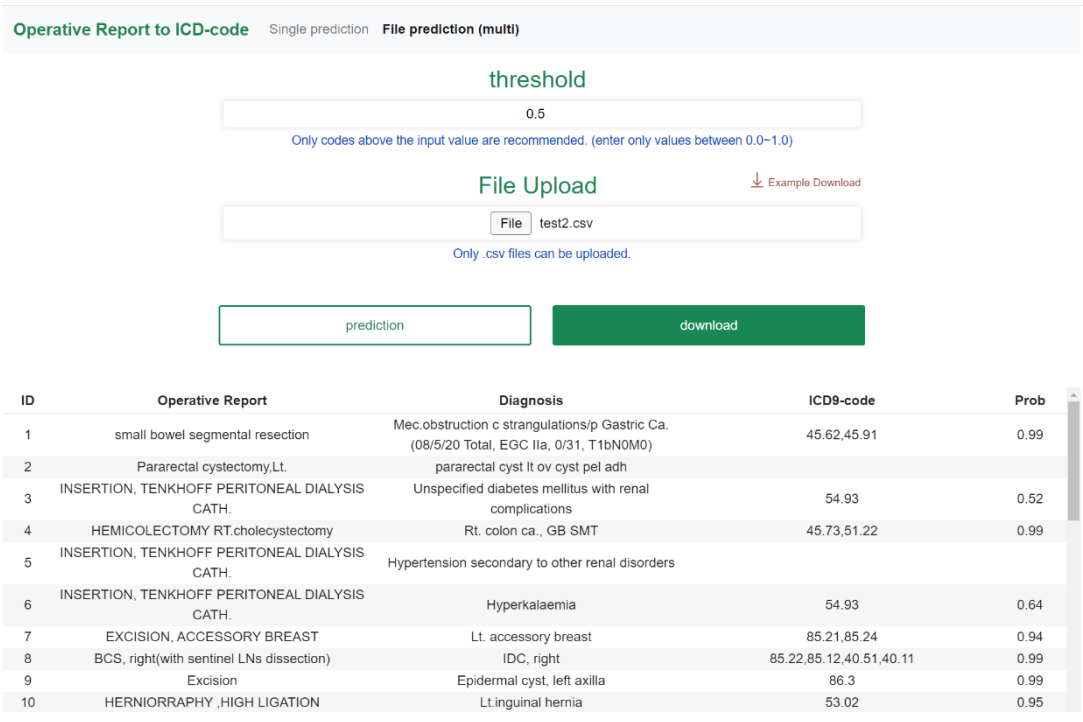


Figure 4. Multi-prediction page of the web-based application.

4. Discussion

In the present study, we developed a ConBERT model using word-and character-level BERT to automate the standardization of operative reports written in free text in electronic medical records with an acceptable prediction capability. Most existing studies have predicted the ICD-9 code with the MIMIC-III dataset using detailed descriptions of diseases such as discharge summaries [25]. In the proposed model, an F1 score of 0.7415 was achieved by predicting the ICD-9 code using only the operative report and diagnosis. Accordingly, by matching various operative reports with ICD-9 codes, surgeons can provide

accurate and standardized data automatically, without modifying the existing methods of recording operative reports in free text. Moreover, manual standardization depends on the quality of the mapping tables or skills, which can differ depending on the individual or the institution. Our model can replace inefficient labor and provide more consistent and accurate standardization of operative reports.

The proposed model utilizes the advantages of both word- and character-level BERT models to solve problems, including non-standardized operative reports, typographical errors, and abbreviations, caused by the data characteristics of surgical records. Thus, compared with the previous studies, the performance was 15–16% higher based on the micro F1 score and 27–29% higher based on the macro F1 score (Table 7).

Table 7. Comparison with existing papers.

Model	F1 Score		AUC	
	Micro	**Macro**	**Micro**	**Macro**
RAC [23]	0.586	0.127	0.991	0.948
Joint LAAT [24]	0.575	0.107	0.988	0.921
Proposed model	0.7415	0.3975	0.9842	0.9703

AUC, area under the curve.

The common data model (CDM), based on common standard terms, was previously proposed as a healthcare data model according to a standardized machine learning framework [26,27]. Observational Health Data Science and Informatics is an international network of researchers and data partners that maintains an analytical solution, the Observational Medical Outcome Partnership (OMOP) CDM, which transforms diverse and heterogeneous databases worldwide into a common format [28,29]. The OMOP CDM enables researchers to investigate standardized content with high consistency and extensibility. The operative report, which contains treatment-related patient information, can be utilized in medical research as an indicator to evaluate the appropriateness of treatments and establish an insurance policy. Therefore, the CDM should include accurate operative reports; therefore, standardization of the operative reports written in free text according to the ICD codes is imperative for building the CDM. Our BERT-based model can serve as an excellent tool to standardize operative reports. Although various tools have been developed to help researchers automatically transform data into standard terminologies [30,31], to the best of our knowledge, a model for standardizing operative reports does not exist. Our model can facilitate a more efficient mapping of operation data to the CDM.

However, our study had several limitations. First, approximately 3% of the data were removed. Although the use of removed data is unlikely in an actual hospital setting, this can hinder accurate automated collection. Therefore, in subsequent studies, it is necessary to collect more data and secure data with a single label to solve this problem. Second, our dataset is highly imbalanced. We conducted learning by giving weights for each class through data-level approaches, but it was confirmed that there is a trade-off in which macro performance increases and micro performance decreases. Further research on data imbalance is needed using weight balancing techniques, or more advanced training loss such as focal loss [32] or self-adjusting dice loss [33]. Third, the study data were limited to patients from the department of surgery. For better performance and predictability, it is necessary to integrate data from multiple institutions and other departments, such as orthopedics and neurosurgery, to expand the scope of our model. Fourth, as ICD codes have recently been revised, a study on converting the ICD-9 codes to the ICD-10 codes should be conducted. Fifth, a commonly used threshold of 0.5 was used in this paper. Performance was evaluated using AP score, which is a precision-recall change measurement method according to confidence-threshold. However, threshold value of 0.5 was used in Web-based applications. The number of appearances of icd-9 codes may vary depending on surgical characteristics. A follow-up study is needed to find the optimal threshold. Despite these limitations, our proposed model showed excellent efficiency in

automating the standardization process of operative reports and an extensional property that can be utilized in the CDM mapping process.

5. Conclusions

We developed a ConBERT model using pre-trained and Character BERT to standardize operative reports as per ICD-9 codes. Our model showed acceptable performance, with a micro AP score of 0.7671, F1 score of 0.7415, and AUC of 0.9842. Using our model, the manual standardization process can be automated, and potential data errors can be reduced. Additionally, our model can be further developed to integrate data from multiple institutions and other departments and may be utilized in mapping operation data to the CDM.

Author Contributions: Conceptualization, S.K., Y.K. and H.-S.C.; methodology, S.K. and H.-S.C.; software, S.P. and I.P.; validation, H.L.; formal analysis, S.P., I.P. and P.P.; data curation, J.C. and J.-W.B.; writing—original draft preparation, S.P. and J.-W.B.; writing—review and editing, S.K., H.L. and H.-S.C.; visualization, S.P.; supervision, S.K. and H.-S.C.; funding acquisition, S.K. and Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea University Guro Hospital (KOREA RESEARCH-DRIVEN HOSPITAL), a grant funded by Korea University Medicine (No. K2018091), and a 2022 Research Grant from Kangwon National University. It was also funded by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1F1A1076454), funded by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2022RIS-005) and by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (20009972, Support for commercialization of next generation semiconductor technology development) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea). And It was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program (IITP-2022-RS-2022-00156439) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation).

Institutional Review Board Statement: Ethical clearances were obtained from the respective Institutional Review Boards of the Korea University Anam Hospital and the Korea University Guro Hospital (2021AN0210, 2020GR0511).

Informed Consent Statement: The requirement for patient consent was waived by the institutional review board of Korea University due to the retrospective nature of this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Safiejko, K.; Tarkowski, R.; Koselak, M.; Juchimiuk, M.; Tarasik, A.; Pruc, M.; Smereka, J.; Szarpak, L. Robotic-assisted vs. standard laparoscopic surgery for rectal cancer resection: A systematic review and meta-analysis of 19,731 patients. *Cancers* **2021**, *14*, 180. [CrossRef] [PubMed]
2. Kim, M.J.; Park, J.W.; Lee, M.A.; Lim, H.K.; Kwon, Y.H.; Ryoo, S.B.; Park, K.J.; Jeong, S.Y. Two dominant patterns of low anterior resection syndrome and their effects on patients' quality of life. *Sci. Rep.* **2021**, *11*, 3538. [CrossRef] [PubMed]
3. Almeida, M.S.C.; Sousa, L.F.; Rabello, P.M.; Santiago, B.M. International Classification of Diseases—11th revision: From design to implementation. *Rev. Saude Publica* **2020**, *54*, 104. [CrossRef] [PubMed]
4. Baumel, T.; Nassour-Kassis, J.; Cohen, R.; Elhadad, M.; Elhadad, N. Multi-label classification of patient notes: Case study on ICD code assignment. *AAAI Workshops* **2018**, arXiv:1709.09587, 409–416.
5. Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; Carin, L. Joint embedding of words and labels for text classification. *arXiv* **2018**, arXiv:1805.04174v1, 2321–2331.
6. Song, C.; Zhang, S.; Sadoughi, N.; Xie, P.; Xing, E. Generalized zero-shot ICD coding. *arXiv* **2019**, arXiv:1909.13154.
7. Haoran, S.; Xie, P.; Hu, Z.; Zhang, M.; Xing, E.P. Towards automated ICD coding using deep learning. *arXiv* **2017**, arXiv:1711.04075.
8. Fei, L.; Hong, Y. ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. In Proceedings of the AAAI, New York, NY, USA, 7–12 February 2020.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
10. Boukkouri, H.E.; Ferret, O.; Lavergne, T.; Noji, H.; Zweigenbaum, P.; Tsujii, J. Character BERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. *arXiv* **2020**, arXiv:2010.10392.

11. NLTK (Natural Language Toolkit). Available online: https://www.nltk.org/ (accessed on 8 September 2022).
12. Heo, T.S.; Yongmin, Y.; Park, Y.; Jo, B.-C. Medical Code Prediction from Discharge Summary: Document to Sequence BERT Using Sequence Attention. In Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, 13–16 December 2021; pp. 1239–1244.
13. Scikit-Learn. Available online: https://scikit-learn.org/stable/ (accessed on 8 September 2022).
14. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
15. Pennington, J.; Socher, R.; Manning, C. GLoVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014.
16. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput Linguist* **2017**, *5*, 135–146. [CrossRef]
17. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1809.09795.
18. Michalopoulos, G.; Wang, Y.; Kaka, H.; Chen, H.; Wong, A. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *arXiv* **2020**, arXiv:2010.10391.
19. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [PubMed]
20. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
21. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of Tricks for Image Classification with Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
22. Pytorch Lightning. Available online: https://www.pytorchlightning.ai/ (accessed on 24 October 2022).
23. Kim, B.H.; Ganapathi, V. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In Proceedings of the Machine Learning for Healthcare Conference, Online, 6–7 August 2021.
24. Vu, T.; Nguyen, D.Q.; Nguyen, A. A Label Attention Model for ICD Coding from Clinical Text. In Proceedings of the IJCAI, Yokohama, Japan, 11–17 July 2020.
25. Johnson, A.E.W.; Pollar, T.J.; Shen, L.; Lehmen, L.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef]
26. Hripcsak, G.; Duke, J.D.; Shah, N.H.; Reich, C.G.; Huser, V.; Schuemie, M.J.; Suchard, M.A.; Park, R.W.; Wong, I.C.K.; Rijnbeek, P.R.; et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **2015**, *216*, 574–578.
27. Ryu, B.; Yoo, S.; Kim, S.; Choi, J. Thirty-day hospital readmission prediction model based on common data model with weather and air quality data. *Sci. Rep.* **2021**, *11*, 23313. [CrossRef]
28. Reps, J.M.; Schuemie, M.J.; Suchard, M.A.; Ryan, P.B.; Rijnbeek, P.R. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 969–975. [CrossRef]
29. Jung, H.; Yoo, S.; Kim, S.; Heo, E.; Kim, B.; Lee, H.Y.; Hwang, H. Patient-level fall risk prediction using the observational medical outcomes partnership's common data model: Pilot feasibility study. *JMIR Med. Inform.* **2022**, *10*, e35104. [CrossRef]
30. Biedermann, P.; Ong, R.; Davydov, A.; Orlova, A.; Solovyev, P.; Sun, H.; Wetherill, G.; Brand, M.; Didden, E.M. Standardizing registry data to the OMOP common data model: Experience from three pulmonary hypertension databases. *BMC Med. Res. Methodol.* **2021**, *21*, 238. [CrossRef] [PubMed]
31. Lamer, A.; Abou-Arab, O.; Bourgeois, A.; Parrot, A.; Popoff, B.; Beuscart, J.B.; Tavernier, B.; Moussa, M.D. Transforming anesthesia data into the observational medical outcomes partnership common data model: Development and usability study. *J. Med. Internet Res.* **2021**, *23*, e29259. [CrossRef] [PubMed]
32. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
33. Self-Adj-Dice. Available online: https://github.com/fursovia/self-adj-dice/ (accessed on 8 September 2022).