Korean Bio-Medical Corpus (KBMC) for Medical Named Entity Recognition

Sungjoo Byun[1], Jiseung Hong[2], Sumin Park[1], Dongjun Jang[1],
Jean Seo[1], Minseok Kim[1], Chaeyoung Oh[1], Hyopil Shin[1]

[1]Seoul National University
{byunsj, mam3b, qwer4107, seemdog, snumin44, nyong10, hpshin}@snu.ac.kr
[2]KAIST
jiseung.hong@kaist.ac.kr

Abstract

Named Entity Recognition (NER) plays a pivotal role in medical Natural Language Processing (NLP). Yet, there has not been an open-source medical NER dataset specifically for the Korean language. To address this, we utilized ChatGPT to assist in constructing the KBMC (Korean Bio-Medical Corpus), which we are now presenting to the public. With the KBMC dataset, we noticed an impressive 20% increase in medical NER performance compared to models trained on general Korean NER datasets. This research underscores the significant benefits and importance of using specialized tools and datasets, like ChatGPT, to enhance language processing in specialized fields such as healthcare.

Keywords: Medical NER, Korean NER dataset, Domain-specific, Data construction with LLM

1. Introduction

The significance of domain-specific Named Entity Recognition (NER), especially in fields like law and medicine, calls for more in-depth research and investigation. The role of NER in medical NLP is as follows: Firstly, NER contributes to processing medical terminology. Medical NER enables language models to identify and process medical terminologies and jargon. Next, it facilitates information extraction from unstructured data. In fact, Pearson et al. (2021) have performed NER to remove or encode information from an unstructured medical dataset. Moreover, NER contributes to entity identification and the anonymization of sensitive patient-specific information (Catelli et al., 2021).

However, it is problematic that medical NER datasets are insufficient. This problem becomes even more challenging as domain-specific NER tasks require extensive labeling, particularly for specific entity categories like Disease, Body, and Treatment. The difficulty is further amplified due to the necessity of expert-level knowledge in medical domains. The data scarcity issue worsens in relatively low-resource languages like Korean. The fact that there is no open-source medical NER dataset for Korean demonstrates the severity of the problem. In order to resolve the data scarcity problem, we introduce KBMC (Korean Bio-Medical Corpus), the first open-source medical NER dataset for Korean. We utilize ChatGPT[1] for effective sentence creation. Subsequently, we annotate entities corresponding to disease name, body part, and treatment following the BIO format. To augment the dataset and to check the performance in general text as well,

we concatenate the Naver dataset,[2] which is the Korean NER dataset with our KBMC in the experiment.

In our research, we evaluate the effectiveness and utility of KBMC by comparing the performance of multiple language models. These models either use a general NER dataset (solely the Naver NER dataset) or a domain-specific dataset (a combination of the Naver NER dataset and KBMC). The results demonstrate that our dataset significantly enhances the accurate recognition of medical entities by more than 20 percent.

Contributions of our research are as follows:

- We describe and publicly release Korean Bio-Medical Named Entity Recognition Corpus (KBMC), the first open-source Korean medical NER dataset. This contributes to solving the data scarcity problem.

- Our research aims to play crucial role in medical data processing. Medical NER would facilitate the sensitive data anonymization process and contribute to the reconstruction of medical data that lack standardized formats.

2. Related Work

Medical NER

As a part of the entity representation task, various studies, mainly in English, have explored the medical field. Traditional research has conducted bio-medical NER using Long short-term memory (LSTM) models (Liu et al., 2017; Lyu et al., 2017;

---

[1]https://chat.openai.com

[2]https://github.com/naver/nlp-challenge

Cho and Lee, 2019). Peng et al. (2019) test Biomedical Language Understanding Evaluation (BLUE) benchmark, including NER with BERT and ELMo. Bio-BERT, a pre-trained language representation model for biomedical text mining, shows high performance in bio-NER (Lee et al., 2019). Also, various toolkits that facilitate clinical NER implemented using SpaCy (Eyre et al., 2021), Apache Spark (Kocaman and Talby, 2022), and Flair (Weber et al., 2021) have been introduced.

Medical NER dataset

Medical NER is crucial, as shown by numerous medical concept extraction challenges, such as those hosted by i2b2 (Uzuner et al., 2011) and n2c2 (Henry et al., 2019). To tackle the data scarcity in the medical field, SemClinBER, a Portuguese medical NER dataset, was introduced by Oliveira et al. (2022), and a Chinese NER dataset was developed by Cheng et al. (2021). Additionally, NCBI-disease (Dogan et al., 2014) and BC5CDR (Li et al., 2016) provide annotations for medical entities in PubMed abstracts. To further address limited data, strategies including data augmentation (Ding et al., 2020), few-shot approaches (Hofer et al., 2018; Yang and Katiyar, 2020; Wang et al., 2021), cross-lingual transfer learning (Chaudhary et al., 2018; Zhou et al., 2022), and web-based annotation tools (Tarcar et al., 2020) have been employed.

3. KBMC : Korean Bio-Medical Corpus

3.1. Data Construction

We use the ChatGPT API[3] to create sentences that include medical terminology such as disease names, body parts, and treatments. Given the availability of comprehensive medical domain knowledge and the capabilities of the large language model, we augment the sentences that include medical terminology via responses from gpt-3.5-turbo. The prompts are designed as "Create a Korean sentence comprising more than 20 words that includes *given medical terminology*". All the sentences augmented by ChatGPT undergo thorough review and verification to mitigate the risk of hallucination issues. Medical terms are downloaded from the Korean Standard Terminology Of Medicine (KOSTOM).[4] It includes 8th revised terms of Korean Standard Classification of Diseases (KCD)[5] and local terms used in the medical field. To facilitate the annotation process, we develop a pre-annotation
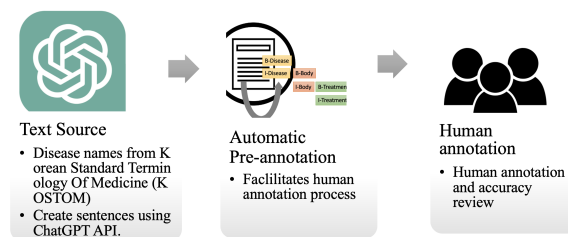
Figure 1: Construction Process of KBMC

Figure 2: KBMC Annotation

algorithm that automatically assigns Named Entity tags as a preliminary step.

Given a set of collected sentences $W = \{w_1, w_2, ..., w_N\}$, we tokenize each sentence using Open-source Korean Text Processor (OKT)[6] so that the input data can be expressed as $\hat{W} = \{x_1, x_2, ..., x_M\}$, where x_i indicates each token. Also, we establish a vocabulary lists for three entity types $E = \{Disease, Body, Treatment\}$. Then, for each entity type $e \in E$, we detect a set of spans $S_e = \{s_{jk} \mid s_{jk} = \{x_j, ..., x_k\} \subset \hat{W}\}$ that matches the vocabulary in the list. Lastly, the algorithm automatically annotates the first token of the span x_j with a **B-** tag, such as **B-Disease**, and annotates the rest of the tokens with an **I-** tag, such as **I-Disease**. In other words, $\forall e \in E, \forall s_{jk} \in S_e$, the pre-annotation of each token x_i can be described as:

$$\text{Annotate}(x_i) = \begin{cases} \textbf{B-}e, \text{for } i = j, \\ \textbf{I-}e, \text{for } j < i \leq k \end{cases}$$

After the pre-annotation process, four human annotators proceed to annotate Named Entities such as "Disease," "Body," and "Treatment" following the BIO format. The annotators modify and correct any incorrect pre-annotations. Subsequently, a separate fifth reviewer reviews the accuracy of the data
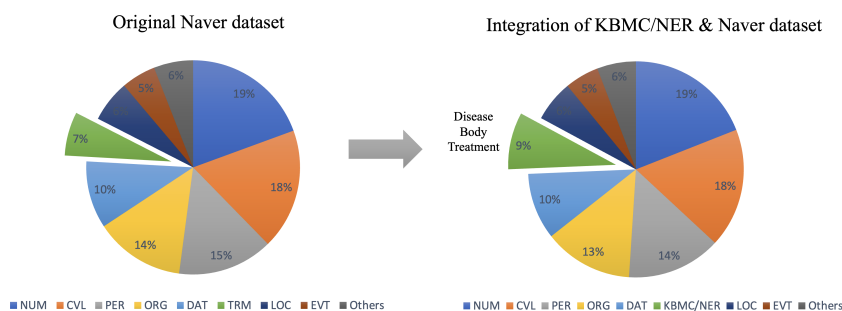
Figure 3: The distribution of Named Entity labels in two datasets: the original Naver NER dataset (left), and a combined version of the Naver NER dataset (partial) and KBMC (right). The original Naver dataset contains the label TRM, representing medical and IT-related terms. In the combined dataset, sentences that include TRM from the original dataset have been replaced with data from KBMC, aiming to achieve a more accurate classification of medical terms into refined categories.

annotation for quality control. The annotation process essentially adheres to the standards of the Korean Standard Terminology of Medicine (KOSTOM). However, if there is a disagreement between the annotator and the third annotator (reviewer) regarding the annotation results, the remaining annotators collectively review the mismatched terminology. Figure Figure 1 summarizes the data construction process. Please refer to Appendix A for example sentences of KBMC.

3.2. Annotation Result

KBMC consists of 6,150 sentences, 153,971 tokens in total. Table 1 displays the label distribution of our dataset. The dataset includes 4,162 distinct disease names, 841 body parts, and 396 treatments. Figure 2 shows the results of the KBMC annotation.

We utilize the OKT (Open-source Korean Text) tokenizer for constructing KBMC. While many Korean NER datasets employ word-level annotation, this approach can be problematic for Korean text. Specifically, word-level tokenization often fails to distinguish between nouns and associated postpositional particles, leading to imprecise annotations by attributing a single Named Entity tag to combined terms and particles. Given that Korean is an agglutinative language, tokenizing at the morpheme level is more precise. Thus, unlike conventional Korean NER datasets, we tokenize sentences into morphemes to ensure more accurate annotations.

3.3. Data Application

For data augmentation and comparison of NER in general and domain-specific text, the Naver NER dataset[7] is concatenated with KBMC. The Naver

Named Entity (NE)	Scheme	# of NE
Disease	B (Begin)	10,595
	I (Inside)	10,089
Body	B (Begin)	5,215
	I (Inside)	1,158
Treatment	B (Begin)	1,193
	I (Inside)	839

Table 1: Label Distribution of KBMC

NER dataset is a general NER dataset, published by Naver[8] and Changwon University. The Naver NER dataset comprises 90,000 sentences and includes 14 named entities, such as PER (Person), FLD (Field), NUM (Number), DAT (Date), and ORG (Organization). Specifically, the dataset includes annotated named entities labeled as TRM (TERM), which refer to medicine and IT-related terminology. To prevent any potential mismatches when concatenated with our KBMC, we exclude 12,426 sentences containing TRM from the Naver NER dataset. The concatenated version of the Naver NER dataset and KBMC includes 13 general Named Entities and 3 medical Named Entities, totaling 16 Named Entities. The integration of the datasets and their label distribution are demonstrated in Figure 3.

4. KBMC

In this section, we compare the performance between the utilization of the Naver dataset, also known as a general NER dataset, and the application of KBMC. Next, to assess the applicability of KBMC, we conduct NER using MedSpaCy[9].

Model	Avg.F1(General)	medical NE	F1 of medical NER
KM-BERT (Kim et al., 2022)	87.08	TRM	75.35
KR-BERT (Lee et al., 2020b)	86.51	TRM	75.26
Ko-BERT	88.01	TRM	78.21
KR-ELECTRA (Lee and Shin, 2022)	87.62	TRM	76.25
Ko-ELECTRA	88.00	TRM	76.58
BiLSTM-CRF (Huang et al., 2015)	55.23	TRM	42.23

Table 2: Medical Named Entities and NER Performance: General NER dataset (The Naver Dataset) solely used.

4.1. Models

The data variation experiment confirms and quantifies the impact of the training data. We evaluate our dataset on six different language models: KM-BERT (Kim et al., 2022), KR-BERT (Lee et al., 2020a), KoBERT[10], KR-ELECTRA[11], Ko-ELECTRA v3[12], and BiLSTM-CRF (Huang et al., 2015). These are advanced Korean NLP models, each with unique architectures and approaches to understanding language. While KR-BERT, KoBERT, KR-ELECTRA, and KoELECTRA v3 leverage transformer-based architectures to achieve state-of-the-art performance on various NLP tasks, BiLSTM-CRF combines bidirectional long short-term memory units with a conditional random field layer, catering to tasks such as NER. Among these models, KM-BERT is a domain-specific language model that has been trained on the Korean medical corpus. Both the learning rate (ranging from 1e-5 to 5e-5) and the batch size (ranging from 32 to 128) are adjusted for optimal performance.

4.2. Results

Medical NER using general dataset

We initially fine-tune six language models using the Naver dataset, which primarily contains general labels. For the experiments, the dataset is split into 90% for training and 10% for testing. All medical entities in this dataset are grouped under one label, TRM. However, this label is not solely for medical terms; it also includes IT-related entities. This generalization makes it difficult to accurately identify and differentiate medical terms since they are consolidated with IT terms under TRM. As a result, the identification of specific medical terminology becomes challenging. Additionally, the F1 score for medical NER using the Naver dataset is below average, as indicated in Table 2.

Model	Avg.F1(General)	Medical NEs	F1 of Medical NER
KM-BERT	88.53 (+1.45)	Disease	98.04 (+22.69)
		Body	98.13 (+22.78)
		Treatment	98.53 (+23.18)
KR-BERT	87.48 (+0.97)	Disease	98.04 (+22.78)
		Body	98.32 (+23.06)
		Treatment	97.82 (+22.56)
KoBERT	88.70 (+0.69)	Disease	98.25 (+20.04)
		Body	98.22 (+20.01)
		Treatment	98.18 (+19.97)
KR-ELECTRA	88.63 (+1.01)	Disease	98.21 (+21.96)
		Body	98.31 (+22.06)
		Treatment	98.53 (+22.28)
KoELECTRA	88.86 (+0.86)	Disease	98.05 (+21.47)
		Body	97.72 (+21.14)
		Treatment	96.56 (+19.98)
BiLSTM-CRF	56.68 (+1.45)	Disease	88.18 (+45.95)
		Body	81.44 (+39.21)
		Treatment	61.14 (+18.91)

Table 3: Medical Named Entities and Performance: KBMC applied. The numbers in blue indicate the degree of improvement when compared to the experimental results in Table 2.

	Avg.F1	Precision	Recall
MedSpaCy	95.69	97.02	95.52

Table 4: Performance of MedSpaCy NER using KBMC

Medical NER using KBMC

We introduce KBMC to address the shortcomings of the Naver dataset. By combining the Naver NER dataset (excluding sentences with TRM) with KBMC, we achieve a more balanced dataset. The average F1 score in Table 3 encompasses 13 general entities from the Naver dataset (TRM excluded) and 3 medical entities from KBMC. The dataset is divided into 90% for training and 10% for testing. To avoid data imbalance, we maintain consistent proportions of general and medical data in both training and testing phases. KBMC offers precise categorization, separating medical entities from IT-related ones and allowing for detailed classification. This specificity results in a performance increase, with the F1 scores for Disease, Body, and Treatment labels surpassing the TRM label by nearly 20 points. The consistent performance across different models demonstrates the quality and reliability of our dataset.

4.3. KBMC Applicability Assessment

Medical NER using MedSpaCy

In order to test the utility of KBMC, we also test our dataset using MedSpaCy. Eyre et al. (2021) have released a library of tools for clinical NLP and text processing with SpaCy.[13] We apply MedSpaCy on Korean dataset by using ko_core_news_md,[14] a

---

[10] https://github.com/SKTBrain/KoBERT
[11] https://github.com/snunlp/KR-ELECTRA
[12] https://github.com/monologg/KoELECTRA

[13] https://spacy.io/
[14] https://spacy.io/models/ko#ko_core_news_md

pretrained statistical model for Korean provided by SpaCy. As shown in Table 4, our KBMC dataset demonstrates remarkable performance on a clinical text processing toolkit in Python as well. While MedSpaCy may not be primed for general entity recognition, it excels in identifying medical terms, especially when enhanced with KBMC.

5. Conclusion

In our research, we introduce KBMC, the first open-source biomedical NER dataset tailored for the Korean language. KBMC provides a training ground for language models to detect and categorize medical Named Entities, addressing the issue of data scarcity in this domain.

We evaluate the utility of the KBMC dataset in two scenarios: one using only a pre-existing general NER dataset, and another incorporating the KBMC dataset. The inclusion of KBMC resulted in enhanced predictions for medical Named Entities and an elevated overall F1 score, which averages the F1 scores for both general and medical entities. With KBMC, models can recognize a broader spectrum of medical terms. Notably, when paired with MedSpaCy, a Python toolkit designed for clinical NLP, our dataset showcases impressive results.

We anticipate that our KBMC dataset will contribute substantially to ongoing research in the field of medical NLP.

Limitations

The primary challenge arises from the limited availability of Korean medical data, which makes it difficult to develop a comprehensive corpus. Due to this constraint, we were unable to manually create a labeled dataset for downstream tasks other than NER task. As a result, an important avenue for future research lies in the construction of a more expansive and diverse Korean medical corpus to facilitate the development of other downstream tasks, such as question-answering (QA). Moreover, while our intention was to compare different general NER datasets in terms of medical entity extraction, The Naver dataset was the only available Korean NER dataset that provided annotations for medical terminology. This kind of problem also occurred in terms of domain-specific models as well. KM-BERT was the only medical language model available for our testing. This limited access to resources restricted our capacity for a comprehensive comparison.

Ethics Statement

Using our KBMC dataset enables precise identification of entity categories. When implemented in the medical sphere, our dataset and model can assist in de-identifying personal details of patients. In the realm of medical NLP, transferring and accessing data is challenging due to the presence of sensitive content. To address these privacy and data sensitivity issues, integrating medical NER into real-world medical institutions offers a safeguarded approach. Resolving these challenges sets the stage for a flourishing future in NLP research, spanning areas such as the medical and legal fields.

Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. A novel covid-19 data set and an effective deep learning approach for the de-identification of italian medical records. *IEEE Access*, PP:1–1.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime G. Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations.

Ming Cheng, Shufeng Xiong, Fei Li, Pan Liang, and Jianbo Gao. 2021. Multi-task learning for chinese clinical named entity recognition with external knowledge. *BMC Medical Informatics and Decision Making*, 21.

Hyejin Cho and Hyunju Lee. 2019. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, 20.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks.

Rezarta Dogan, Robert Leaman, and Zhiyong lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo J. Nevado-Holgado. 2018. Few-shot learning for named entity recognition in medical text. *CoRR*, abs/1811.05468.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Yoojoong Kim, Jeong Lee, Moon Jang, Yun Yum, Seongtae Kim, Unsub Shin, Young-Min Kim, Hyung Joo, and Sanghoun Song. 2022. A pre-trained bert for korean medical natural language processing. *Scientific Reports*, 12:13847.

Veysel Kocaman and David Talby. 2022. Accurate clinical and biomedical named entity recognition at scale. *Software Impacts*, 13:100373.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020a. Kr-bert: A small-scale korean-specific language model.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020b. Kr-bert: A small-scale korean-specific language model.

Sangah Lee and Hyopil Shin. 2022. Kr-electra: a korean-based electra model. https://github.com/snunlp/KR-ELECTRA.

Jiao Li, Yueping Sun, Robin Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn Mattingly, Thomas Wiegers, and Zhiyong lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Zengjian Liu, Ming Yang, Xiaolong Wang, Qingcai Chen, Buzhou Tang, Zhe Wang, and Wang Qi. 2017. Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making*, 17.

Chen Lyu, Bo Chen, Yafeng Ren, and Donghong Ji. 2017. Long short-term memory rnn for biomedical named entity recognition. *BMC Bioinformatics*, 18.

Lucas Emanuel Silva Oliveira, Ana Carolina Peters, Adalniza Moura Pucca da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1).

Cole Pearson, Naeem Seliya, and Rushit Dave. 2021. Named entity recognition in unstructured medical text documents.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

Amogh Kamat Tarcar, Aashis Tiwari, Vineet Naique Dhaimodker, Penjo Rebelo, Rahul Desai, and Dattaraj Rao. 2020. Healthcare ner models using language model pretraining.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Yaqing Wang, Haoda Chu, Chao Zhang, and Jing Gao. 2021. Learning from language description: Low-shot named entity recognition via decomposed framework.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. Conner: Consistency training for cross-lingual named entity recognition.

Appendix

A. Appendix

KBMC sentences	Translation	NER Tags
전신 적 다한증 은 신체 전체 에 힘 이 빠져서 일상 생활 이 어려 워지는 질환 으로 , 근육 통증 과 무기 력 감 이 동반 됩니다 .	Systemic myasthenia is a condition in which the whole body loses strength, making daily life difficult, accompanied by muscle pain and a sense of lethargy.	Disease-B Disease-I Disease-I O O O O O O O O O O O O O Disease-B Disease-I O Disease-B Disease-I Disease-I O O O O
췌장암 이란 췌장 에 생긴 암세 포 로 이루어진 종괴 (종양 덩어 리) 이다 .	Pancreatic cancer refers to a tumor (a lump of tumor) made up of cancer cells that form in the pancreas.	Disease-B O Body-B O O O O O Disease-B O Disease-B O O O O
이러한 병명 은 폐 기능 저하 로 인한 호흡 곤란 기침 천식 발작 등 의 증상 을 유발 하 여 일상생 활 에 큰 영향 을 미칩니다 .	Such diseases lead to symptoms such as respiratory distress, coughing, asthma attacks, etc., caused by decreased lung function, greatly affecting daily life.	O O O Disease-B Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I O O O O O O O O O O O O O O
버킷 림프종 은 림프절 에서 발 생 하는 악성 종양 으로 , 조기 발 견 과 치료 가 중요하며 항암 치 료 나 방사선 치료 등 다양한 치 료법 이 존재 합니다 .	Burkitt lymphoma is a malignant tumor that originates in the lymph nodes. Early detection and treatment are crucial, and various treatment methods, such as chemotherapy and radiation therapy, exist.	Disease-B Disease-I O Body-B O O O Disease-B Disease-I O O O O O O O O Treatment-B Treatment-I O Treatment-B Treatment-I O O O O O O O

Table 5: Examples of the KBMC dataset