Self-Supervised Intermediate Fine-Tuning of Biomedical Language Models for Interpreting Patient Case Descriptions

Israa Alghanmi[*], **Luis Espinosa-Anke**[*◇], **Steven Schockaert**[*]
[*]CardiffNLP, Cardiff University, UK
[◇]AMPLYFI, UK
{alghanmiia, espinosa-ankel, schockaerts1}@cardiff.ac.uk

Abstract

Interpreting patient case descriptions has emerged as a challenging problem for biomedical NLP, where the aim is typically to predict diagnoses, to recommended treatments, or to answer questions about cases more generally. Previous work has found that biomedical language models often lack the knowledge that is needed for such tasks. In this paper, we aim to improve their performance through a self-supervised intermediate fine-tuning strategy based on PubMed abstracts. Our solution builds on the observation that many of these abstracts are case reports, and thus essentially patient case descriptions. As a general strategy, we propose to fine-tune biomedical language models on the task of predicting masked medical concepts from such abstracts. We find that the success of this strategy crucially depends on the selection of the medical concepts to be masked. By ensuring that these concepts are sufficiently salient, we can substantially boost the performance of biomedical language models, achieving state-of-the-art results on two benchmarks.

1 Introduction

Natural Language Processing (NLP) in the biomedical domain poses a number of particular challenges. For this reason, several Language Models (LMs) that are specialised towards the biomedical domain have been proposed, including BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), SciBERT (Beltagy et al., 2019a), and PubMedBERT (Gu et al., 2021). Recent work has focused on analysing the capabilities of such models (Jin et al., 2019; Alghanmi et al., 2021; Sung et al., 2021) and enhancing them further (He et al., 2020b; Yuan et al., 2021; Zhang et al., 2021; Fei et al., 2021). Broadly speaking, biomedical LMs have proven successful in capturing the meaning of specialised terminology, but they have been far less successful in enabling medical reasoning, e.g. for predicting a likely diagnosis from a given patient case description. Currently, the main strategies for alleviating this latter issue have centered on incorporating structured knowledge, especially in the form of knowledge graphs. For example, Meng et al. (2021) proposed a method to integrate a large biomedical knowledge graph into a language model through the use of adapters (Pfeiffer et al., 2020), while Zhang et al. (2022) used graph neural networks to jointly reason about language model outputs and knowledge graphs.

Question: A 38-year-old woman comes to the emergency department because of progressive headache, blurry vision, and nausea for 1 day. Four days ago, she was diagnosed with a right middle ear infection. She appears lethargic. Her temperature is 39.1°C (102.3°F), and blood pressure is 148/95 mm Hg. Ophthalmologic examination shows bilateral swelling of the optic disc. The corneal reflex in the right eye is absent. Sensation to touch is reduced on the upper right side of the face. Serum studies show increased concentrations of fibrin degradation products. Which of the following is the most likely diagnosis?
(A) Cerebral venous thrombosis
(B) Hypertensive emergency
(C) Subarachnoid hemorrhage
(D) Viral meningitis

Table 1: Example of a question from MedQA, along with the answer candidates.

We focus on the task of interpreting patient case descriptions. To illustrate this task, Table 1 shows a question from the MedQA benchmark (Jin et al., 2021). In this context, the aim is typically to infer a diagnosis or to recommend a treatment. This is highly challenging, even for biomedical language models, because many pieces of information may need to be combined to find the right answer, and often some degree of clinical judgment is needed. Accordingly, the performance of state-of-the-art biomedical language models remains rather low for benchmarks such as MedQA. We argue that this can, to some extent, be explained by the fact that

Proceedings of the 29th International Conference on Computational Linguistics, pages 1432–1441
October 12–17, 2022.

interpreting patient case descriptions is a paragraph-level task, whereas the standard masked language modelling objective encourages the model to primarily focus on sentence-level context.

Ideally, biomedical language models for interpreting patient case descriptions would be pre-trained on a task that involves predicting diagnoses, or other salient aspects of these patient cases. Unfortunately, beyond the training fragment of benchmarks such as MedQA, such labelled data is not readily available. As an alternative, we propose to generate a pseudo-labelled dataset, based on the heuristic that whenever a case descriptions mentions a disease, it is likely (although by no means guaranteed) that this disease is a valid diagnosis, and similar for other medical concepts such as treatments. To get access to a large set of case descriptions, we rely on abstracts of published case reports. In particular, starting from a collection of PubMed abstracts, we first use a simple heuristic to identify those that are likely to correspond to case reports. Given a case report that mentions some disease, we then fine-tune the language model on the task of predicting that disease. Note that the target disease is masked, as the task would otherwise be trivial. The pre-training task is formulated as a binary classification problem, i.e. given a patient description and a disease, is that disease the correct diagnosis (or more precisely, is it the disease that was masked). This formulation has the advantage that the input format is similar to that of multiple-choice question answering (QA) and natural language inference (NLI). Beyond diseases, we also experiment with predicting masked treatments. Similar to the usual masked language modelling objective, our pre-training task involves making predictions about masked text spans. However, due to the fact that we specifically mask diseases and treatments, we hypothesize that this will improve the model's ability to take the whole case description into account when making predictions. Finally, note that we consider this to be an intermediate fine-tuning step. In other words, we start from a state-of-the-art biomedical language model, which is then fine-tuned on the proposed task, before finally being fine-tuned on a downstream task.

We find that this intermediate fine-tuning leads to substantial improvements in downstream tasks, even when using a biomedical LM that was already pre-trained on PubMed. To some extent, this comes from the fact that we specifically fine-tune the model on case reports. However, this in itself is not sufficient. To achieve good results, we find that a careful selection of the target concepts is needed. For instance, strong results are obtained when only masking medical treatments. When masking diseases, the improvements over the baseline are sometimes smaller. This is surprising, given that most questions in the considered benchmarks are about diagnosing diseases. Upon closer inspection, the under-performance of strategies that rely on masking diseases appears to be related to the fact that diseases can be mentioned for two common reasons: (i) because the patient has been diagnosed with that disease, which is the case that underpins the intuition behind our proposed approach, or (ii) because the disease is relevant to the medical history of the patient. In the latter case, only a small part of the abstract may be relevant to the disease, which hampers the extent to which the model learns to focus on the case description as a whole. To address this issue, we propose to split abstracts in which multiple diseases are mentioned. Despite the simplicity of the overall approach, our fine-tuning strategies enable significant improvements over the current state-of-the-art in two benchmarks that are focused on patient case descriptions: MedQA (Jin et al., 2021) and DisKnE (Alghanmi et al., 2021). [1]

2 Related Work

The standard paradigm in NLP at the moment is to fine-tune a pre-trained LM, such as BERT (Devlin et al., 2018), on task-specific training data. However, it has been observed that adding an intermediate step, where the LM is first fine-tuned on a different task, for which training data is more abundant, can be highly beneficial (Phang et al., 2018, 2020; Oğuz et al., 2021; Park and Caragea, 2020; Poth et al., 2021). Several works have investigated the role of intermediate tasks, in particular with the aim of analyzing when and why results improve (Pruksachatkun et al., 2020; Chang and Lu, 2021).

For the biomedical domain, one strategy has been to rely on transfer learning from general-domain tasks. For instance, Soni and Roberts (2020) use general-domain question answering for intermediate training, to improve a clinical question answering system. Another strategy has been to rely on different, but related tasks, such as pre-

[1] Code and data are available at: https://github.com/israa-alghanmi/Intermediate-FT-Biomedical-LMs

training on natural language inference to develop a question answering system (Jeong et al., 2020). Furthermore, several authors have proposed techniques for infusing the knowledge from biomedical knowledge graphs into LMs (He et al., 2020a; Meng et al., 2021; Jha and Zhang, 2022). More closely related to our approach, He et al. (2020c) propose a strategy which relies on the structure of Wikipedia to infuse knowledge about diseases. For instance, to teach the model about how diseases are treated, they rely on the fact that disease-centric Wikipedia articles tend to have a section called *Treatment*. They then combine the content of that section with a generated question-style sentence mentioning the aspect considered (i.e. treatment in this case) and a masked disease. However, rather than infusing encylopedic knowledge, our aim is to teach LMs to interpret patient case descriptions. Another related approach was introduced by Pergola et al. (2021), who propose to fine-tune a biomedical language model by using a masked language modelling objective which is modified such that only biomedical concepts are masked. This approach has some similarities with our work, e.g. the idea of masking biomedical concepts as an intermediate fine-tuning task, but there are also some clear differences. First, we formulate our task as a binary classification problem, rather than masked language modelling. Moreover, we specifically target diseases and treatments, and we only mask one concept at a time (although all occurrences of that concept are masked). Finally, since we focus on paragraph-level understanding, we pay particular attention to how these input paragraphs can be selected. As we will see, each of these differences has a clear impact on the empirical results.

3 Proposed Method

We consider the problem of making predictions from patient case descriptions. For instance, given a description that lists symptoms and other information about the patient (e.g. gender, age, and medical history), we would like to infer the corresponding diagnosis or to recommend suitable treatments. We are specifically interested in the potential of using freely available case reports from the medical literature to improve the ability of standard biomedical LMs to make such predictions. In Section 3.1, we first explain our overall strategy. Subsequently, in Section 3.2 we describe the specific variants that we included in our analysis.

3.1 Overall Strategy

Our aim is to design an intermediate fine-tuning task for specialising biomedical LMs towards the task of interpreting patient case descriptions. This fine-tuning task relies on passages from PubMedDS (Vashishth et al., 2021), a corpus which primarily consists of abstracts from PubMed. First, we split the abstracts into passages of up to 250 words, to address the limitations on input length of BERT-based LMs. Next, we aim to identify those passages that contain a case report, describing a specific patient rather than more general findings. To this end, we rely on the simple but effective heuristic that case reports often mention the age of the patient. In particular, we select those passages that contain at least one keyword from the following list: *year-old male*, *year-old female*, *year-old boy*, *year-old girl*, *year-old woman*, *year-old man*. Let us write $\mathcal{D}$ for the resulting corpus, i.e. the set of passages that contain at least one of the aforementioned keywords. Subsequently, we determine which medical concepts are mentioned in the passages from $\mathcal{D}$. To this end, we use QuickUMLS (Soldaini and Goharian, 2016) with UMLS-2020AA to identify both the spans and the semantic types (e.g. diseases, treatments) of the mentioned concepts. Finally, we create positive training examples of the form (P,C), where C is a medical concept, and P is a passage from $\mathcal{D}$ in which all mentions of C have been replaced by a single <*mask*> token. To generate negative training examples, we simply replace the medical concept C by another concept, as explained below. A given example (*passage*,*concept*) is encoded as follows: "<*cls*> *passage* <*sep*> *concept*", mimicking the input format that is typically used for question answering and natural language inference models. The LM is fine-tuned on these examples using a standard cross-entropy loss.

3.2 Training Strategies

We now describe the different variants that we considered. These variants primarily differ in the kinds of medical concepts that are selected as target concepts. Across all variants, we never mask the concept *disorder*, as constructing training examples from such mentions was found to be highly detrimental, given its prevalence and generic meaning. For all variants, we attempt to balance the number of positive and negative examples. Table 2 provides an overview of the total number of training exam-

	#
AnyType	1,011,482
SpecificType	
– *diseases*	160,534
– *treatments*	2,460
SplitDis	100,225
OneDis	3,310

Table 2: The total number of training examples for each of the intermediate fine-tuning tasks (#).

ples arising from each of the following strategies.

AnyType We create a positive example for every medical concept that is found (with the exception of *disorder*). Note that passages typically mention several concepts, hence this strategy allows us to derive multiple positive examples from the same passage, each time masking a different concept. To construct negative examples, we corrupt positive examples by randomly selecting a concept from those that have been identified in the corpus, regardless of the semantic type.

SpecificType In this variant, we only construct training examples from medical concepts of particular types. Specifically, we have experimented with diseases and treatments. Negative examples are constructed by replacing the target concept with another concept of the same semantic type, i.e. diseases are replaced by diseases, and treatments are replaced by treatments.

SplitDis Many passages contain more than one disease, which may confuse the model. For instance, diseases which are mentioned as part of the patient history may only be loosely related to the rest of the case report. Since our aim is to train the model to make predictions based on the whole case description, in this variant, passages containing more than one disease are split into sub-passages. In particular, when constructing a positive example for a target disease d, we select the sub-passage which begins with the first sentence in which d is mentioned, and includes all the subsequent sentences, until we reach a sentence that mentions another disease (where this final sentence is excluded from the selected sub-passage). If the target disease is mentioned in a sentence that also contains another disease, it is excluded altogether. For illustration, training examples that were obtained with the *SplitDis* strategy are presented in Table 3.

OneDis Instead of splitting passages mentioning more than one disease into sub-passages, as with SplitDis, here we simply discard such passages. This results in a much smaller number of positive examples, but with stronger guarantees that the disease being masked is salient. In both this and the *SplitDis* method, negative examples are obtained by using randomly selected diseases.

4 Experiments

In this section, we empirically analyse the different variants of the intermediate fine-tuning strategy.

Evaluation Datasets We mainly focus on two benchmarks that are specifically focused on interpreting and reasoning about patient cases. First, we use **MedQA** (Jin et al., 2021), which is a multiple-choice question answering benchmark. The questions are taken from medical exams and are specifically asking about what can be inferred from a given patient case description. We use the English version of this dataset (USMLE). Results for this benchmark are reported in terms of accuracy (Acc). Second, we use **DisKnE** (Alghanmi et al., 2021), which has been derived from MedNLI (Romanov and Shivade, 2018). Therefore, to use DisKnE, a license and access to MedNLI is required. Instances of this benchmark consist of a patient case description and a disease, and the aim is to predict whether that disease can be inferred as diagnosis. This repurposing from the original MedNLI is of particular relevance to our experiments, given that many instances in MedNLI can be solved simply with linguistic knowledge. DisKnE contains a separate training-test split for each disease, and for each split, we consider the task of ranking all test cases, according to our confidence that the given target disease is a valid diagnosis. The results are averaged across all diseases and are reported in terms of Mean Average Precision (MAP). We use the medical-similar variant of the benchmark.

In addition, we also consider the English version of **HeadQA** (Vilares and Gómez-Rodríguez, 2019), as a more general healthcare-oriented QA dataset. This dataset contains a broad variety of healthcare questions, most of which do not involve patient descriptions. However, the questions are designed to require complex medical reasoning. As such, we use this benchmark to analyse whether our proposed approach may also benefit such settings.

SpecificType-treatments	The role of [MASK] in the treatment of a patient with a pure silent pituitary somatotroph carcinoma. To describe a case of a pure silent somatotroph pituitary carcinoma. We describe a 54-year-old female with a clinically nonfunctioning pituitary macroadenoma diagnosed 15 years earlier. The patient underwent transsphenoidal surgery and no visible tumor remnant was observed for 6 years. A magnetic resonance imaging (MRI) detected the recurrence of a 1.2 × 1.5 cm macroadenoma. The patient was submitted to conventional radiotherapy (4500 cGy), and the tumor volume remained stable for 7 years. Then, an MRI revealed a slight increase in tumor size, and 2 years later, a subsequent MRI detected a very large, invasive pituitary mass. The patient was resubmitted to transsphenoidal surgery, and the histopathological examination showed diffuse positivity for growth hormone (GH). The nadir GH level during an oral glucose tolerance test was 0. 06 ng/mL, and the pre- and postoperative insulin like growth factor type I (IGF-I) levels were within the normal range. Abdominal, chest, brain, and spine MRI showed multiple small and hypervascular liver and bone lesions suggestive of metastases. Liver biopsy confirmed metastasis of GH-producing pituitary carcinoma. The patient has been treated with [MASK] and zoledronic acid for 7 months and with octreotide long-acting release (LAR) for 4 months. → *Temozolomide*
SpecificType-diseases	Intestinal cholesterol absorption inhibitor ezetimibe added to cholestyramine for sitosterolemia and xanthomatosis. Sitosterolemia is a rare, recessively inherited disorder characterized by increased absorption and delayed removal of noncholesterol sterols, which is associated with accelerated atherosclerosis, premature [MASK], hemolysis, and xanthomatosis. Treatments include low-sterol diet and bile salt-binding resins; however, these often do not reduce the xanthomatosis. We examined the effects of the intestinal cholesterol/phytosterol transporter inhibitor ezetimibe added to cholestyramine in a young female patient with sitosterolemia and associated xanthomatosis. The patient was an 11-year-old female with sitosterolemia presenting with prominent xanthomas in the subcutaneous tissue of both elbows who was receiving treatment with cholestyramine 2 g once daily. Bilateral carotid bruits were audible, and a grade II/VI systolic murmur was detected at the left upper sternal border. She also had a low platelet count of 111,000/microL. Ezetimibe 10 mg once daily was added to the patient's ongoing cholestyramine regimen, and she was evaluated for 1 year. The patient followed an unrestricted diet during the 1-year treatment period. After 1 year of treatment with ezetimibe added to ongoing cholestyramine therapy, the patient's plasma sitosterol and campesterol levels decreased by approximately 50. → *coronary artery disease*
SplitDis	After initial improvement artificial ventilation had to be be gun on day 3 because of an acute [MASK], diagnosed both clinically and radiologically. Despite additional antiviral and intensive medical treatment he died on day 11. → *respiratory distress syndrome*
	Traumatic [MASK] present diagnostic and therapeutic challenges. Owing to their fragile nature, endovascular intervention has become the first-line treatment; however, direct surgery has an advantage in certain cases. → *intracranial aneurysms*
	A fluoroscopic sniff test demonstrated diaphragmatic dysfunction and pulmonary function tests revealed [MASK] with evidence of neuromuscular etiology. → *restrictive pulmonary disease*

Table 3: Examples obtained with the different variants of the proposed strategies.

Some questions in this dataset require interpreting images. As this is beyond the scope of the paper, we discard all questions involving images for our experiments. This resulted in a total number of 2589 questions for training, 1336 for validation, and 2675 for testing.

Setup We use four pre-trained LMs for the baselines and main experiments:

- the cased version of the standard $\text{BERT}_{\text{base}}$ (Devlin et al., 2019);
- the cased version of SciBERT (Beltagy et al., 2019b);
- the cased version of ClinicalBERT (Alsentzer et al., 2019) that was trained on MIMIC-III while being initialized from BioBERT (Lee et al., 2020);
- the PubMedBERT model (Gu et al., 2021) that was trained from scratch on full-length PubMed articles as well as abstracts.

As a baseline, we directly fine-tune the models on the training data from the downstream task. For the other configurations, we first fine-tune the models on the proposed intermediate task.

We use the official training, validation, and test splits for each dataset, with the exception that we excluded questions with images for HeadQA.

Training Details We use the same settings and hyper-parameters for all datasets. For fine-tuning the models on the target task, we set the batch size to 8, the number of epochs to 4 and the learning rate to 2e-5. For the intermediate fine-tuning step, we again set the batch size to 8 and the learning rate to 2e-5. Regarding the number of epochs for intermediate fine-tuning, we note that the number of training examples varies greatly across the different variants. For this reason, and to mitigate the potential for catastrophic forgetting, we tuned the number of epochs, choosing from $\{2, 3, 4\}$, based on the development split of the downstream task.

Limitations Our method relies on an automated extraction tool for identifying the target medical concepts, which will inevitably lead to some noisy training examples. For example, *SplitDis* and *OneDis* rely on the assumption that we can detect all mentions of diseases in the text. More generally, regardless of performance, the predictions of a biomedical LM can clearly not be relied upon for diagnosing patients or recommending treatments in a clinical setting. Our purpose in studying these models is rather because a deeper understanding of patient records would make it possible to improve retrieval systems (e.g. suggesting relevant case reports to a clinician handling an unusual patient) or to identify hypotheses for medical research (e.g. by inducing patterns from large sets of case reports).

4.1 Results

Tables 4, 5 and 6 summarize our results. As can be seen, PubMedBERT clearly outperforms the other language models. In general, most variants of the intermediate fine-tuning tasks lead to clear improvements over the baselines. A clear and remarkable conclusion that can be observed for all benchmarks is that the type of intermediate fine-tuning data appears to be much more important than the number of training examples. For instance, the version of *SpecificType* which only uses treatments achieves the best overall results, outperforming the previous state-of-the-art for MedQA and achieving among the strongest results for both DisKnE and HeadQA. This is surprising, both because of the small number of training examples we can generate for this variant and because of the focus on diseases in DisKnE and many of the MedQA and HeadQA questions.

For MedQA, *SpecificType* with treatments outperforms the previous state-of-the-art (Zhang et al., 2022) by 1.9 percentage points, despite not relying on any structured knowledge graphs. Note that DisKnE is a recent benchmark, for which the only reported results thus far were obtained from simply fine-tuning biomedical LMs. These existing results were reported prior to the introduction of PubMedBERT, which outperforms these published results. The *OneDis* variant performs well for DisKnE, despite the low number of corresponding training examples. For MedQA, *SplitDis* outperforms *SpecificType* with diseases (with the exception of BERT), which supports the idea that simply masking diseases can lead to training examples that are too noisy. While HeadQA is not particularly focused on patients case descriptions, we still see consistent improvements over the baselines with *SpecificType*, *SplitDis* and *OneDis*, although the improvements are somewhat smaller than those for MedQA and DisKnE.

We can see that our proposed strategy outperforms the baselines for each of the different language models, with the exception of SciBERT with DisKnE. However, there are some differences between the language models in terms of which variant of our method performs best. For MedQA, for instance, we can see that *SpecificType* with diseases is highly competitive for BERT and ClinicalBERT (compared to the other variants for these language models). As these are the language models that are least adapted to the considered task, we can indeed expect that more pre-training data might be needed for these models. This can explain the relative success of *SpecificType* with diseases and *SplitDis*, given that these are associated with a larger number of training examples.

	BERT	ClinicalBERT	SciBERT	PubMedBERT
Baseline	27.8	29.1	29.2	35.5
AnyType	**28.2**	31.2	32.7	36.5
SpecificType				
– *diseases*	**28.2**	31.5	30.4	38.0
– *treatments*	27.8	31.0	**34.5**	**40.4**
SplitDis	27.7	**31.8**	33.4	38.7
OneDis	27.0	29.6	33.3	35.6

Table 4: Results for MedQA in terms of Accuracy.

	BERT	ClinicalBERT	SciBERT	PubMedBERT
Baseline	57.0	67.5	**69.2**	69.7
AnyType	**64.2**	71.6	68.8	71.9
SpecificType				
– *diseases*	60.2	70.0	67.0	72.9
– *treatments*	57.5	67.5	68.3	73.6
SplitDis	58.3	**74.1**	68.1	72.2
OneDis	64.0	68.2	66.2	**74.4**

Table 5: Results for DisKnE in terms of Mean Average Precision (MAP).

4.2 Analysis

Table 7 shows the results of some variants of the *SpecificType* with diseases and *SplitDis* strategies, as explained next. We use PubMedBERT for these experiments, as this model achieved the best results in the main experiments. We focus on the MedQA benchmark as this is the most representative benchmark for our problem setting.

Frequent vs Rare We analyze whether there is any advantage in focusing specifically on common diseases, or conversely, in focusing on rare diseases. Table 7 shows the results of two variants of Speci-

	BERT	ClinicalBERT	SciBERT	PubMedBERT
Baseline	28.8	29.3	32.8	39.5
AnyType	29.3	30.0	31.7	39.1
SpecificType				
– *diseases*	29.8	30.1	34.5	**41.8**
– *treatments*	**30.3**	**31.1**	**35.7**	41.0
SplitDis	29.8	29.6	32.6	40.7
OneDis	29.7	29.8	34.0	40.8

Table 6: Results for HeadQA in terms of Accuracy.

		#	**MedQA** (Acc)
SpecificType	Most-Frequent	49,816	36.8
	Least-Frequent	8,466	38.0
	Most-General	7,229	36.6
	Most-Specific	8,778	36.9
	Treatment-Case-Dis	6,934	38.2
SplitDis	Most-Similar	1,858	37.7
	Least-Similar	1,870	36.7
	SplitDis+Def	105,952	37.7
	Treatment-Case-Dis	2,430	38.4

Table 7: Analysis results for MedQA (Accuracy). We also report the total number of training examples for each of the intermediate fine-tuning tasks (#). Results were obtained using PubMedBERT.

ficType, called *Most-Frequent* and *Least-Frequent*. The former only considers training examples, for the intermediate fine-tuning task, involving the 50 diseases which are most common in our corpus of case reports. Similarly, the *Least-Frequent* variant only considers the 5000 least frequent diseases. *Least-Frequent* achieves the best result, despite involving far fewer training examples than *Most-Frequent*. The results of both variants are either below or similar to those with the full set of diseases in Table 4.

General vs Specific Rather than selecting diseases based on their number of occurrences, here we investigate the effect of choosing diseases based on whether they are general or specific, in terms of the level at which they appear in the SNOMED CT hierarchies (Stearns et al., 2001). Specifically, for the *Most-General* variant, we only consider diseases with fewer than 5 ancestors in SNOMED CT. For the *Most-Specific* variant, we only consider diseases with at least 30 ancestors. We find that both variants of *SpecificType* perform similarly.

Similar vs Different We explore a setting in which only case reports about diseases similar to "heart disease" are provided during training. Specifically, we use cui2vec (Beam et al., 2020) to identify the 50 most similar diseases that occur at least once in our corpus of case reports. We then consider a variant of *SplitDis* in which only passages with *heart disease*, or any of the 50 similar diseases, occurs as the target disease. Our aim in this experiment is to see whether training on one type of diseases is sufficient to obtain good results. Furthermore, we may also assume that because the resulting corpus only involves similar diseases, the model is forced to focus on more subtle details in the paragraphs, and might thus improve as a result. To test this hypothesis, we also consider the variant *Least-Similar*, where we instead use the diseases that are least similar to *heart disease*. Rather than fixing the number of diseases at 50, in this case we chose the number to ensure a similar number of training examples as for *Most-Similar*. The results for both variants are below those of the standard *SplitDis* variant. However, we can see that *Most-Similar* clearly outperforms *Least-Similar*.

Adding Definitions We analyse the usefulness of UMLS definitions. Specifically, we augment the SplitDis training examples with examples of the form (def, dis), where def is the UMLS definition of a disease, and dis is the corresponding disease. Negative examples are again created by replacing the target disease with a randomly chosen other disease. The results in Table 7 show that adding definitions does not improve the results.

Diseases in Treatment Cases The good performance of the SpecificType variant with treatments, despite the small number of training examples we have for that setting, is one of the most surprising findings from the main experiments. Here we analyse whether this might be related to the quality of the case reports that were selected in that setting, i.e. the case reports that mention a treatment. To this end, we consider all such case reports, but instead of using the treatments as the target concepts, we instead focus on diseases. In other words, we use the SpecificType setting for diseases, but applied to

	MedQA (Acc)
MLM-RandomMask	
– *SplitDis*	36.4
– *SpecificType: treatments*	35.2
MLM-SpecificMask	
– *SplitDis*	37.6
– *SpecificType: treatments*	38.5
Random-Abstracts	
– *SplitDis*	38.2
– *SpecificType: treatments*	37.6
No Mask	
– *SplitDis*	36.7
– *SpecificType: treatments*	37.8
Remove-Sent (treatments)	38.9

Table 8: Ablation results for MedQA in terms of Accuracy. Results were obtained using PubMedBERT.

the case reports that mention treatments. We also consider a variant in which the *SplitDis* setting is applied to these case reports. The results in Table 8, shown as *Treatment-Case-Dis*, reveal that this variant still underperforms the *SpecificCase* variant with treatments.

4.3 Ablation Experiments

In this section, we analyse the importance of a number of our design choices. We again focus on PubMedBERT and MedQA. We specifically consider the SplitDis and SpecificType with treatments, as these yielded the best results in the main experiments. The results are summarized in Table 8.

Masked Language Modelling We experimented with two variants of the masked language modelling (MLM) objective for the intermediate fine-tuning task. For the *MLM-RandomMask* variant, we randomly mask tokens, following the standard approach that is used for LM pre-training. For the *MLM-SpecificMask* variant, we specifically mask the tokens corresponding to diseases (for the SplitDis setting) or treatments (for the SpecificType setting). The results show that our approach outperforms both MLM strategies, while *MLM-SpecificMask* outperforms *MLM-RandomMask*.

Random Abstracts vs Case Reports We analyse the importance of specifically focusing on case reports. In the *Random-Abstracts* variant, rather than targeting abstracts which are likely to correspond to case reports, we consider a set of 60,000 randomly sampled abstracts from PubMedDS. We then use our SplitDis and SpecificType settings to construct the examples. The results in Table 8 show that using randomly chosen abstracts leads to worse results, compared to our standard setting.

Masking vs not Masking We consider a variant of the method in which the original passage is used, i.e. where we do not replace occurrences of the target disease with a <*mask*> token. The results in Table 8 clearly shows that masking is essential to achieve the best results. Nonetheless, even without masking we obtain results that are clearly better than those of the baseline (i.e. PubMedBERT without intermediate fine-tuning).

Masking vs Removing Sentences Instead of replacing the target concept with a <*mask*> token, here we remove the entire sentence in which this concept is mentioned. For this variant, called *Remove-Sent*, we only consider the SpecificType setting (with treatments), as using SplitDis would result in too few examples, given that several SplitDis examples consist of a single sentence. The results show that removing the sentence underperforms masking the concept.

5 Conclusions

We have proposed a strategy for intermediate fine-tuning of biomedical language models, to improve their ability to interpret patient case descriptions. The core of our strategy is to exploit abstracts of case reports found in the literature, as a surrogate of patient case descriptions, and to rely on the heuristic that diseases and treatments that are mentioned in such abstracts are likely to correspond to diagnoses and recommendations, respectively. Despite its conceptual simplicity and without the cost of manual annotation, this approach was found to lead to clear performance gains, setting a new state-of-the-art in MedQA and DisKnE, while also benefitting more diverse datasets such as HeadQA.

References

Israa Alghanmi, Luis Espinosa Anke, and Steven Schockaert. 2021. Probing pre-trained language models for disease knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3023–3033, Online. Association for Computational Linguistics.

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

Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. 2020. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing*, volume 25, pages 295–306.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019a. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3613–3618.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019b. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Ting-Yun Chang and Chi-Jen Lu. 2021. Rethinking why intermediate-task fine-tuning works. *arXiv preprint arXiv:2108.11696*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3):bbaa110.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020a. Bert-mk: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020b. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020c. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4604–4614.

Minbyul Jeong, Mujeen Sung, Gangwoo Kim, Donghyeon Kim, Wonjin Yoon, Jaehyo Yoo, and Jaewoo Kang. 2020. Transferability of natural language inference to biomedical question answering. *arXiv preprint arXiv:2007.00217*.

Kishlay Jha and Aidong Zhang. 2022. Continual knowledge infusion into pre-trained biomedical language models. *Bioinformatics*, 38(2):494–502.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zaiqiao Meng, Fangyu Liu, Thomas Hikaru Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-partitions: Infusing large biomedical knowledge graphs into bert. *arXiv preprint arXiv:2109.04810*.

Barlas Oğuz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen-tau Yih, Sonal Gupta, et al. 2021. Domain-matched pre-training tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*.

Seoyeon Park and Cornelia Caragea. 2020. Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5409–5419, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. 2021. Boosting low-resource biomedical qa via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1977–1985.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Sarvesh Soni and Kirk Roberts. 2020. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5532–5538.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734.

Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics*, 121:103880.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190.

Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. 2021. SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5882–5893, Online. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. GreaseLM: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.