Probing Pre-Trained Language Models for Disease Knowledge

Israa Alghanmi, Luis Espinosa-Anke, Steven Schockaert
Cardiff University, United Kingdom
{alghanmiia,espinosa-ankel,schockaerts1}@cardiff.ac.uk

Abstract

Pre-trained language models such as ClinicalBERT have achieved impressive results on tasks such as medical Natural Language Inference. At first glance, this may suggest that these models are able to perform medical reasoning tasks, such as mapping symptoms to diseases. However, we find that standard benchmarks such as MedNLI contain relatively few examples that require such forms of reasoning. To better understand the medical reasoning capabilities of existing language models, in this paper we introduce DisKnE, a new benchmark for Disease Knowledge Evaluation. To construct this benchmark, we annotated each positive MedNLI example with the types of medical reasoning that are needed. We then created negative examples by corrupting these positive examples in an adversarial way. Furthermore, we define training-test splits per disease, ensuring that no knowledge about test diseases can be learned from the training data, and we canonicalize the formulation of the hypotheses to avoid the presence of artefacts. This leads to a number of binary classification problems, one for each type of reasoning and each disease. When analysing pre-trained models for the clinical/biomedical domain on the proposed benchmark, we find that their performance drops considerably.

1 Introduction

Pre-trained language models (LMs) such as BERT (Devlin et al., 2019) are currently the de-facto architecture for solving most NLP tasks, and their prevalence in general language understanding tasks is today indisputable (Wang et al., 2018, 2019). Beyond generic benchmarks, it has been shown that LMs are also extremely powerful in domain-specific NLP tasks, e.g., in the biomedical domain (Lewis et al., 2020). While there are several reasons why they are preferred over standard neural architectures, one important (and perhaps less obvious) reason is that LMs capture a substantial amount of world knowledge. For instance, several authors have found that LMs are able to answer questions without having access to external resources (Petroni et al., 2019; Roberts et al., 2020), or that they exhibit commonsense knowledge (Forbes et al., 2019; Davison et al., 2019). To analyze the capabilities of LMs in a more systematic way, there is a growing interest in designing probing tasks, which are now common across the NLP landscape, e.g., for word and sentence-level semantics (Paperno et al., 2016; Conneau et al., 2018). In this paper we focus on (generic and specialized) LMs in the biomedical domain, and ask the following question: what kinds of medical knowledge do pre-trained LMs capture? More specifically, we focus on *disease knowledge*, which encompasses for instance the ability to link symptoms to diseases, or treatments to diseases.

Among the several biomedical LMs (i.e. LMs that have been pre-trained on biomedical text corpora) that exist today, some of the most prominent are SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019). Rather than architectural features, these models differ from each other mostly in the pre-training corpora: SciBERT was trained from scratch on scientific papers; BioBERT is an adapted version of BERT (Devlin et al., 2019), which was fine-tuned on PubMed articles as well as some full text biomedical articles; and ClinicalBERT was initialized from BioBERT and further fine-tuned on MIMIC-III notes (Johnson et al., 2016), which are clinical notes describing patients admitted to critical care units. These LMs have enabled impressive results on various reading comprehension benchmarks for the medical domain, such as MedNLI (Romanov and Shivade, 2018) and MEDIQA-NLI (Abacha et al., 2019) for Natural Language Infer-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3023–3033

ence (NLI), and PubMedQA (Jin et al., 2019b) for QA. As an example, Wu et al. (2019) achieved an accuracy of 98% on MEDIQA-NLI, which might suggest that medical NLI is essentially a solved problem. This would be exciting, as medical NLI intuitively requires a wealth of medical knowledge, much of which is not available in structured form.

However, a closer inspection of MedNLI, the most well-known medical NLI benchmark, reveals three important limitations, namely: (1) only few test instances actually require *medical disease* knowledge, with instances that (only) require terminological and lexical knowledge (e.g. understanding acronyms or paraphrases) being more prevalent; (2) training and test examples often cover the same diseases, and thus it cannot be determined whether good performance comes from the capabilities of the pre-trained LM itself, or from the fact that the model can exploit similarities between training and test examples; and (3) hypothesis-only baselines perform rather well on MedNLI, which shows that this benchmark has artefacts that can be exploited, similarly to general-purpose NLI benchmarks (Poliak et al., 2018).

We therefore propose DisKnE (Disease Knowledge Evaluation), a new benchmark for evaluating biomedical LMs. This dataset explicitly addresses the three limitations listed above and thus constitutes a more reliable testbed for evaluating the disease knowledge captured by biomedical LMs. DisKnE is derived from MedNLI and is organized into two top-level categories, which cover instances requiring medical and terminological knowledge respectively. The medical category is furthermore divided into four sub-categories, depending on the type of medical knowledge that is required.

We empirically analyse the performance of existing biomedical LMs, as well as the standard BERT model, on the proposed benchmark. Our results show that all the considered LMs struggle with NLI examples that require medical knowledge. We also find that the relative performance of the pre-trained models differs across medical categories, where the best performance is obtained by ClinicalBERT, BioBERT, SciBERT or BERT depending on the category and experimental setting. Conversely, for examples that are based on terminological knowledge, overall performance is much higher, with relatively little difference between different pre-trained models. The contributions of this paper are as follows[1]:

- We introduce a new benchmark to assess the disease-centred knowledge captured by pre-trained LMs, organised into categories that reflect the type of reasoning that is needed, and with training-test splits that avoid leakage of disease knowledge.
- We analyze the performance of several clinical/biomedical BERT variants on each of the considered categories. We find that all considered models struggle with examples that require medical disease knowledge.
- We find that without canonicalizing the hypotheses, hypothesis-only baselines achieve the best results in some categories. This shows that the original MedNLI dataset suffers from annotation artefacts, even within the set of entailment examples.

2 Related Work & Background

Knowledge Encoded in LMs There is a rapidly growing body of work that is focused on analyzing what knowledge is captured by pre-trained LMs. A recurring challenge in such analyses is to separate the knowledge that is already captured by a pre-trained model from the knowledge that it may acquire during a task-specific fine-tuning step. A common solution to address this is to focus on zero-shot performance, i.e. to focus on tasks that require no fine-tuning, such as filling in a blank (Davison et al., 2019; Talmor et al., 2020). As an alternative strategy, Talmor et al. (2020) propose to analyse the performance of models that were fine-tuned on a small training set. Other work has focused on extracting structured knowledge from pre-trained LMs. Early approaches involved manually designing suitable prompts for extracting particular types of relations (Petroni et al., 2019). Recently, however, several authors have proposed strategies that automatically construct such prompts (Bouraoui et al., 2020; Jiang et al., 2020; Shin et al., 2020). Finally, Bosselut et al. (2019) proposed to fine-tune LMs on knowledge graph triples, with the aim of then using the model to generate new triples.

[1]All code for reconstructing the dataset and replicating the experiments is available at: https://github.com/israa-alghanmi/DisKnE. License and access to MedNLI, MEDIQA-NLI and UMLS will be needed.

LMs for Biomedical Text As already mentioned in the introduction, a number of pre-trained LMs have been released for the biomedical domain. Several authors have analyzed the performance of these models, and the impact of including different types of biomedical corpora in particular. For instance, Peng et al. (2019) proposed an evaluation framework for biomedical language understanding (BLUE). They obtained the best results with a BERT model that was pre-trained on PubMed abstracts and MIMIC-III clinical notes. Another large-scale evaluation of biomedical LMs has been carried out by Lewis et al. (2020). To evaluate the biomedical knowledge that is captured in pre-trained LMs, as opposed to acquired during training, Jin et al. (2019a) freeze the transformer layers during training. They find that when biomedical LMs are thus used as fixed feature extractors, BioELMo outperforms BioBERT. Most closely related to our work, He et al. (2020) recently also highlighted the limited ways in which biomedical LMs capture disease knowledge. To address this, they proposed a pre-training objective which relies on a weak supervision signal, derived from the structure of Wikipedia articles about diseases. Other authors have suggested to include structured knowledge, e.g. from UMLS, during the pre-training stage of BERT-based models (Michalopoulos et al., 2020; Hao et al., 2020). Another strategy is to inject external knowledge into task-specific models (rather than at the pre-training stage), for instance in the form of definitions (Lu et al., 2019) or again UMLS (Sharma et al., 2019). Kearns et al. (2019) presented a related approach to our work in which they categorize each sentence pair according to the tense and focus (e.g. medication, diseases, procedures, location) of the hypothesis, with the aim of providing a detailed examination of MEDIQA-NLI. Based on this categorization, they compare the performance of Enhanced Sequential Inference Model (ESIM) using ClinicalBERT, Embeddings of Semantic Predications (ESP), and cui2vec. However, their analysis was limited to the MEDIAQ-NLI test set, whereas we include entailment examples from the entire MedNLI and MEDIQA-NLI datasets. Moreover, we focus specifically on the ability of LMs to distinguish between closely related diseases, and we move away from the NLI setting to avoid training-test leakage and artefacts.

Adversarial NLI Several Natural Language Inference (NLI) benchmarks have been found to contain artefacts that can be exploited by NLP systems to perform well without actually solving the intended task (Poliak et al., 2018; Gururangan et al., 2018). In particular, it has been found that strong results can often be achieved by only looking at the hypothesis of a (premise, hypothesis) pair. In response to this finding, several strategies for creating harder NLI benchmarks have been proposed. One established approach is to create adversarial stress tests (Naik et al., 2018; Glockner et al., 2018; Aspillaga et al., 2020), in which synthetically generated examples are created to specifically test for phenomena that are known to confuse NLI models. This may, for instance, involve the use of WordNet to obtain nearly identical premise and hypothesis sentences, in which one word is replaced by an antonym or co-hyponym. In this paper, we rely on a somewhat similar strategy, using UMLS to replace diseases in hypotheses. As another strategy to obtain hard NLI datasets, Nie et al. (2020) used human annotators to iteratively construct examples that are incorrectly labelled by a strong baseline model. While the aforementioned works are concerned with open-domain NLI, some work on creating adversarial datasets for the biomedical domain has also been carried out. In particular, Araujo et al. (2020) studied the robustness of systems for biomedical named entity recognition and semantic text similarity, by introducing misspellings and swapping disease names by synonyms. To the best of our knowledge, no adversarial NLI datasets for the biomedical domain have yet been proposed.

3 Dataset Construction

In this section, we describe the process we followed for constructing DisKnE. As we explain in more detail in Section 3.1, this process involved filtering the entailment instances from the MedNLI and MEDIQA-NLI datasets, to select those in which the hypothesis expresses that the patient has (or is likely to have) a particular target disease. These instances were then manually categorized based on the type of knowledge that is needed for recognizing the validity of the entailment. Section 3.2 discusses our strategy for generating negative examples, which were obtained in an adversarial way, by replacing diseases occurring in entailment examples with similar ones. Details of the resulting training-test splits are provided in Section 3.3. In a

Category	# inst.	Premise	Hypothesis
Symptoms → Disease	112	The patient developed neck pain while training with increasing substernal heaviness and left arm pain together with sweating.	The patient has symptoms of acute coronary syndrome
Treatments → Disease	60	The patient started on Mucinex and Robitussin.	The patient has sinus disease
Tests → Disease	116	Cardiac enzymes recorded CK 363, CK-MB 33, TropI 6.78	The patient has cardiac ischemia
		A large R hemisphere ICH was revealed when the patent had head CT	The patient has an aneurysm
Procedures → Disease	70	Bloody fluid was removed by pericardiocentesis	The patient has hemopericardium.
Terminological	259	The patient has urinary tract infection	The patient has a UTI
		The patient has high blood pressure	Hypertension
		Transfusions in the past could be the cause of the patient having hepatitis C	The patient has hepatitis C

Table 1: Considered categories of disease-focused entailment pairs.

final step, we canonicalize the hypotheses of all examples, as explained in Section 3.4. Note that the benchmark we propose consists of binary classification problems (i.e. predicting entailment or not), rather than the standard ternary NLI setting (i.e. predicting entailment, neutral, or contradiction), which is motivated by the fact that natural contradiction examples are hard to find when focusing on disease knowledge.

3.1 Selecting Entailment Pairs

We started from the set of all entailment pairs (i.e. premise-hypothesis pairs labelled with the *entailment* category) from the full MedNLI and MEDIQA-NLI datasets. We used MetaMap to find those pairs whose hypothesis mentions the name of a disease, and to retrieve the UMLS CUI (Concept Unique Identifier) code corresponding to that disease. We then manually identified those pairs, among the ones whose hypothesis mentions a disease, in which the hypothesis specifically expresses that the patient has that disease. For instance, in this step, a number of instances were removed in which the hypothesis expresses that the patient does not have the disease. The remaining cases were manually assigned to categories that reflect the type of disease knowledge that is needed to identify that the hypothesis is entailed by the premise. The considered categories are described in Table 1, which also shows the number of (positive) examples we obtained and illustrative examples[2]. The primary distinction we make is between examples that need medical knowledge and those that need terminological knowledge. The former category is divided into four sub-categories, depending on the type of inference that is needed. First, we have the *symptoms-to-disease* category, containing examples where the premise describes the signs or symptoms exhibited by the patient, and the hypothesis mentions the corresponding diagnosis. Second, we have the *treatments-to-disease* category, where the premise instead describe medications (or other treatments followed by the patient). The third category, *tests-to-disease*, involves instances where the premise describes lab tests and diagnostic tools such as X-rays, CT scans and MRI. Finally, the *procedures-to-disease* category has instances where the premise describes surgeries and therapeutic procedures that the patient underwent. In the *terminological* category, the disease is mentioned in both the premise and hypothesis, either as an abbreviation, a synonym or within a rephrased sentence.

3.2 Generating Examples

The process outlined in Section 3.1 only provides us with positive examples. Unfortunately, MedNLI and MEDIQA-NLI contain only few negative examples (i.e. instances of the *neutral* or *contradiction* categories) in which the hypothesis expresses that the patient has some disease. For this reason, rather than selecting negative examples from these datasets, we generate negative examples by corrupting the positive examples. In particular, to generate negative examples, we replace the disease

[2]For data protection reasons, we only provide synthetic examples, which are different from but similar in spirit to those from the original MedNLI dataset.

X from a given positive example by other diseases $Y_1, ..., Y_n$ that are similar to X, but not ancestors or descendants of X in SNOMED CT (Donnelly et al., 2006). To identify similar diseases, we have relied on cui2vec (Beam et al., 2020), a pre-trained clinical concept embedding that was learned from a combination of insurance claims, clinical notes and biomedical journal articles. Apart from the requirement that the diseases $Y_1, ..., Y_n$ should be similar to X, it is also important that they are sufficiently common diseases, as including unusual diseases would make the corresponding negative examples too easy to detect. For this reason, we only consider the diseases that occur in the hypothesis of other positive examples as candidates for the negative examples. Specifically, among these set of candidate diseases, we selected the $n = 10$ most similar ones to X, which were not descendants or ancestors of X in SNOMED CT (as ancestors and descendants would not necessarily invalidate the entailment). This resulted in a total of 4133 examples requiring medical knowledge and 2639 examples requiring terminological knowledge.

3.3 Training-Test Splits

Because our focus is on evaluating the knowledge captured by pre-trained language models, we want to avoid overlap in the set of diseases in the training and test splits. In other words, if the model is able to correctly identify positive examples for a target disease X, this should be a reflection of the knowledge about X in the pre-trained model, rather than knowledge that it acquired during training. However, any single split into training and test diseases would leave us with a relatively small dataset. For this reason, we consider each disease X in isolation. Let $\mathcal{E}$ be the set of all positive examples, obtained using the process from Section 3.1. Furthermore, we write $\mathcal{E}_X$ for the set of those examples from $\mathcal{E}$ in which the target disease in the hypothesis is X. Finally, we write $neg(X)$ for the set $\{Y_1, ..., Y_n\}$ of associated diseases that was selected to construct negative examples, following the process from Section 3.2.

For each target disease X, we define a corresponding test set $Test_X$ and training set $Train_X$ as follows. $Test_X$ contains all the positive examples from $\mathcal{E}_X$. Moreover, for each $e \in \mathcal{E}_X$ and each $Y \in neg(X)$ we add a negative example $e_{X \to Y}$ to $Test_X$ which is obtained by replacing the occurrence of X by Y. If the word before the occurrence

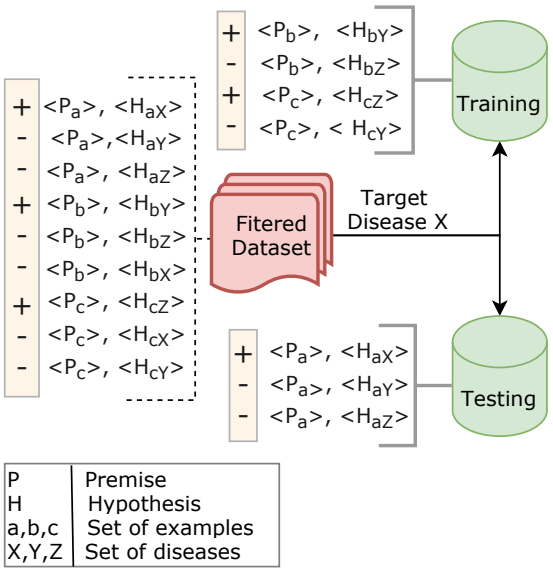

Figure 1: Illustration of training-test splitting process.

of X is *a* or *an*, we modify it depending on whether Y starts with a vowel or consonant. The positive examples in $Train_X$ consist of all examples from $\mathcal{E}$ in which X is not mentioned. Note that we also remove examples in which these diseases are only mentioned in the premise. Furthermore, we check for occurrences of all the synonyms of these diseases that are listed in UMLS. The process of creating the training and test set for a given target disease X is illustrated in Figure 1.

3.4 Canonicalization

We noticed that the way in which a given hypothesis expresses that "the patient has disease X" is correlated with the type of the disease. For this reason, as a final step, we canonicalize the hypotheses in the dataset. Specifically, we replace each hypothesis by the name of the corresponding disease X. Several hypotheses in the dataset already have this form. By converting the other hypotheses in this format, we eliminate any artefacts that are present in their specific formulation.

4 Experiments

We experimentally compare a number of pre-trained biomedical LMs on our proposed DisKnE benchmark. In Section 4.1, we first describe the considered LMs and the experimental setup. The main results are subsequently presented in Section 4.2. This is followed by a discussion in Section 4.3.

	ClinicalBERT	BioBERT	SciBERT	BERT
coronary atherosclerosis	0	0	29	10
chf	67	67	67	67
acs	04	33	0	05
stroke	80	56	90	90
heart disease	80	87	93	100
myocardial infarction	0	0	19	0
heart failure	0	0	22	0
urinary tract infection	100	100	67	100
disorder of lung	89	97	97	100
cirrhosis of liver	0	11	0	0
hyperglycemic disorder	27	13	22	0
pneumonia	89	93	67	100
neurological disease	67	67	80	67
respiratory failure	87	70	22	43
pulmonary edema	74	25	0	50
ami	0	0	0	0
deep vein thrombosis	47	48	50	48
acute cardiac ischemia	0	45	17	72
uri	78	45	67	83
cholangitis	22	22	33	22
atherosclerosis	66	0	67	0
Macro-average	$46_{\pm 3.0}$	$42_{\pm 7.3}$	$43_{\pm 3.1}$	$46_{\pm 3.4}$
Weighted average	$49_{\pm 3.1}$	$47_{\pm 6.0}$	$49_{\pm 2.7}$	$51_{\pm 2.7}$

Table 2: Results for the *Symptoms* → *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

	ClinicalBERT	BioBERT	SciBERT	BERT
chf	55	55	53	55
acs	12	19	0	0
hypertensive disorder	55	67	54	22
heart disease	45	22	0	89
urinary tract infection	100	100	100	100
disorder of lung	82	89	100	93
hyperglycemic disorder	100	69	87	69
pneumonia	60	67	78	57
anemia	17	17	45	22
renal insufficiency	69	89	67	72
pulmonary infection	82	77	89	83
copd	45	67	61	39
hyperlipidemia	59	61	61	55
Macro-average	$60_{\pm 6.1}$	$61_{\pm 1.4}$	$61_{\pm 3.8}$	$58_{\pm 1.6}$
Weighted average	$51_{\pm 5.3}$	$54_{\pm 1.6}$	$51_{\pm 1.7}$	$45_{\pm 2.4}$

Table 3: Results for the *Treatments* → *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

4.1 Experimental Setup

Pre-trained LMs. To understand to what extent the pretraining data of an LM affects its performance on our fine-grained evaluation of disease knowledge, we used the following BERT variants:

BERT We use the $\text{BERT}_{\text{base}}$-cased model (Devlin et al., 2019).

BioBERT Lee et al. (2019) proposed a model based on $\text{BERT}_{\text{base}}$-cased, which they further trained on biomedical corpora. We use the version where PubMed and PMC were utilized for this further pre-training.

ClinicalBERT Alsentzer et al. (2019) introduced four BERT model variants, trained on various clinical corpora. We use the version that was initialized from BioBERT and trained on MIMIC-III notes afterwards.

SciBERT Beltagy et al. (2019) introduced a BERT model variant that was trained from scratch on approximately 1.14M scientific papers from semantic scholar, 82% of which were biomedical articles. The full text of the papers was used for training. We use the cased version.

Training Details. For fine-tuning, model hyperparameters were the same across all BERT variants such as the random seeds, batch size and the learning rate. In this study, we fix the the learning rate at 2e-5, batch size of 8 and we set the maximum number of epochs to 8 with the use of early stopping. We used 10% of the training set as validation split.

Evaluation Protocol. We analyze the results per disease and per category in terms of F1 score for the positive class, reporting results for all diseases that have at least two positive examples for the considered category. To this end, for each disease X, we start from its corresponding training-test split, which was constructed as explained in Section 3.3. To show the results for a particular category, we remove from the test set all the examples that do not belong to that category.

4.2 Results

The main results are shown in Tables 2–6. A number of clear observations can be made. First, the results for the terminological category are substantially higher than the results for the other categories, which suggests that the masked language modelling

	ClinicalBERT	BioBERT	SciBERT	BERT
coronary atherosclerosis	0	0	0	0
chf	52	55	52	55
acs	0	22	0	0
stroke	87	87	95	77
hypertensive disorder	09	26	45	21
myocardial infarction	28	0	30	14
heart failure	0	55	40	0
urinary tract infection	87	90	59	90
hyperglycemic disorder	81	10	68	33
pneumonia	100	100	89	89
anemia	0	0	24	0
aortic valve stenosis	11	24	0	27
syst. inflam. resp. syndr.	76	64	80	80
acute renal failure syndr.	0	0	0	22
chronic renal insufficiency	0	0	0	0
kidney disease	22	0	45	0
ischemia	93	100	93	100
Macro-average	38 ± 2.4	37 ± 1.6	42 ± 3.1	36 ± 5.0
Weighted average	31 ± 2.6	32 ± 1.2	37 ± 1.5	31 ± 3.7

Table 4: Results for the *Tests* $\rightarrow$ *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

	ClinicalBERT	BioBERT	SciBERT	BERT
coronary atherosclerosis	0	0	16	0
heart disease	83	74	84	84
heart failure	33	33	50	0
cirrhosis of liver	0	0	0	0
end stage renal disease	37	29	70	79
respiratory failure	58	27	57	27
renal insufficiency	100	100	93	100
cardiac arrest	100	100	93	100
disorder of resp. syst.	76	80	80	71
peripheral vascular dis.	0	0	78	0
Macro-average	49 ± 3.2	44 ± 5.9	62 ± 3.9	46 ± 5.0
Weighted average	40 ± 3.3	36 ± 7.4	55 ± 5.6	44 ± 4.6

Table 5: Results for the *Procedures* $\rightarrow$ *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

objective, which is used as the main pre-training task in all the considered LMs, may not be ideally suited for learning medical knowledge. Second, recall that the main difference between the considered biomedical LMs comes from the corpora that were used for pre-training them. As the results for the terminological category (Table 6) reveal, the inclusion of domain-specific corpora does not seem to benefit their ability to model biomedical terminology, as similar results for this category are obtained with the standard BERT model, which was pre-trained on Wikipedia and a corpus of books and movie scripts. For the *Symptoms* $\rightarrow$ *Disease* category, we see that ClinicalBERT outperforms the other biomedical LMs, although the standard BERT model actually achieves the best performance overall. The results suggest that ClinicalBERT is better at distinguishing between relatively rare diseases, but that the focus on encyclopedic text benefits BERT for more common diseases. Intuitively, we can indeed expect that the encyclopedic style of Wikipedia focuses more on symptoms of diseases than scientific articles, which might focus more on treatments, procedures and diagnostic tests. This is also in accordance with the findings from He et al. (2020), who obtained promising results with a disease-centric LM pre-training task that relies on Wikipedia. On the *Procedures* $\rightarrow$ *Disease* and *Tests* $\rightarrow$ *Disease* categories, we can see that SciBERT achieves the best results, with a particularly wide margin on the *Procedures* $\rightarrow$ *Disease* category. Finally, for the *Treatments* $\rightarrow$ *Disease* category, the relatively poor performance of BERT stands out, which conforms with the aforementioned intuition that scientific articles put more emphasis on procedures, treatments and tests. BioBERT achieves the best results, although the performance of the other biomedical LMs is quite similar.

4.3 Discussion

Which LM model? Several published works have found ClinicalBERT to outperform the other considered biomedical LMs on biomedical NLP tasks (Alsentzer et al., 2019; Kearns et al., 2019; Hao et al., 2020). In our results, however, SciBERT achieves the most consistent performance, clearly outperforming ClinicalBERT on the *Procedures* $\rightarrow$ *Disease* and *Test* $\rightarrow$ *Disease* categories, while performing similar to ClinicalBERT on the remaining categories. However, rather than providing a blanket recommendation for SciBERT, our fine-grained analysis highlights the fact that different models have different strengths. The most surprising finding, in this respect, is the performance of the standard BERT model, which achieves the best results on the *Symptoms* $\rightarrow$ *Disease* category and

	ClinicalBERT	BioBERT	SciBERT	BERT
anemia	95	100	100	93
aortic valve stenosis	100	100	93	100
carotid artery stenosis	50	50	60	50
coronary atherosclerosis	79	79	76	79
type 2 diabetes mellitus	67	56	64	61
gerd	0	0	0	0
cardiac arrest	95	97	92	97
heart disease	100	100	93	80
heart failure	100	100	100	100
chf	19	37	35	36
hyperglycemic disorder	57	63	80	57
hypertensive disorder	84	87	90	84
acute renal failure synd.	67	67	58	61
end-stage renal disease	77	77	78	70
disorder of lung	89	76	70	52
copd	100	100	97	100
myocardial infarction	24	25	25	21
pancreatitis	33	0	22	33
pleural effusion	80	100	100	80
pneumonia	89	93	89	66
pulmonary edema	87	82	56	76
stroke	81	100	71	100
urinary tract infection	78	77	78	77
aaa	100	96	100	100
Macro-average	73 ± 2.7	73 ± 0.4	72 ± 2.5	70 ± 3.2
Weighted average	74 ± 1.8	76 ± 1.4	75 ± 1.3	72 ± 3.0

Table 6: Results for the terminological category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

performs comparably to BioBERT on several other categories (with *Treatments* $\rightarrow$ *Disease* being a notable exception).

Dataset Artefacts. As already reported by Romanov and Shivade (2018), the original MedNLI dataset has a number of annotation artefacts, which mean that hypothesis-only baselines can perform well. In our dataset, we tried to address this by only using entailment examples, and creating negative examples by corrupting these. However, without canonicalizing the hypotheses, we found that hypothesis-only baselines were still performing rather well. This is shown in Table 7, which summarizes the results we obtained for a version of our dataset without canonicalization, i.e. where the full hypotheses are provided, and the canonicalized version, where the hypotheses were replaced by the disease name only. The table shows results for the standard ClinicalBERT model, as well as for a hypothesis-only variant, which is only given the hypothesis. As can be seen, without canonicalization, the hypothesis only baseline performs similarly to the full model, even outperforming it in a few cases, with the exception of the *Terminological* category where a clear drop in performance for the hypothesis-only baseline can be seen. In contrast, for the canonicalized version of the dataset, we can see that the hypothesis only baseline, which only gets access to the name of the disease in this case, under-performs consistently and substantially. Note that the hypothesis-only baseline still achieves a non-trivial performance in most cases, noting that an uninformed classifier that always predicts true would achieve an F1 score of 0.167. However, this simply shows that the model has learned to prefer

		Standard		Hyp. only	
		full	can	full	can
MACRO	*Symptoms* $\rightarrow$ *Dis.*	48 ± 0.7	46 ± 3.0	47 ± 4.9	23 ± 0.5
	Treatments $\rightarrow$ *Dis.*	64 ± 4.7	60 ± 6.1	65 ± 2.5	29 ± 2.1
	Tests $\rightarrow$ *Dis.*	41 ± 1.7	38 ± 2.4	44 ± 2.3	18 ± 2.0
	Procedures $\rightarrow$ *Dis.*	59 ± 4.9	49 ± 3.2	52 ± 2.6	19 ± 3.0
	Terminological	71 ± 2.3	73 ± 2.7	39 ± 1.3	25 ± 0.4
WEIGHTED	*Symptoms* $\rightarrow$ *Dis.*	54 ± 2.9	49 ± 3.1	53 ± 4.7	23 ± 1.3
	Treatments $\rightarrow$ *Dis.*	62 ± 2.8	51 ± 5.3	60 ± 7.1	24 ± 1.0
	Tests $\rightarrow$ *Dis.*	37 ± 1.4	31 ± 2.6	42 ± 0.2	17 ± 2.8
	Procedures $\rightarrow$ *Dis.*	54 ± 6.2	40 ± 3.3	59 ± 5.1	14 ± 2.0
	Terminological	71 ± 1.1	74 ± 1.8	41 ± 2.7	22 ± 0.4

Table 7: Comparison between a variant with the full hypothesis and the proposed canonicalized version. Results are for the ClinicalBERT model in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

		ClinicalBERT	BioBERT	SciBERT	BERT
MACRO	*Symptoms* $\rightarrow$ *Dis.*	66 ± 4.0	56 ± 3.2	57 ± 5.2	56 ± 4.1
	Treatments $\rightarrow$ *Dis.*	69 ± 4.3	70 ± 2.0	76 ± 4.5	55 ± 4.8
	Tests $\rightarrow$ *Dis.*	53 ± 0.9	49 ± 3.3	52 ± 1.0	47 ± 0.6
	Procedures $\rightarrow$ *Dis.*	60 ± 1.8	56 ± 0.8	76 ± 2.6	60 ± 4.5
	Terminological	77 ± 0.9	77 ± 0.6	74 ± 0.6	76 ± 1.0
WEIGHTED	*Symptoms* $\rightarrow$ *Dis.*	66 ± 5.2	59 ± 3.5	59 ± 4.1	56 ± 4.6
	Treatments $\rightarrow$ *Dis.*	64 ± 6.2	59 ± 3.6	68 ± 4.8	46 ± 3.1
	Tests $\rightarrow$ *Dis.*	53 ± 0.6	51 ± 2.4	54 ± 1.6	43 ± 4.0
	Procedures $\rightarrow$ *Dis.*	65 ± 3.0	58 ± 1.0	76 ± 0.4	67 ± 4.5
	Terminological	76 ± 1.6	77 ± 1.0	75 ± 0.4	72 ± 0.7

Table 8: Results for a variant of our benchmark, in which negative examples were selected at random, in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

frequent diseases over rare ones.

Adversarial Examples. A key design choice has been to select negative examples from the diseases that are most similar to the target disease. To analyse the impact of this choice, we carried out an experiment in which negative examples were instead randomly selected. As before, we only consider diseases that are present in the dataset, and we ensure that negative examples are not ancestors or descendants of the target disease in SNOMED CT. The results are presented in Table 8. As expected, the results are overall higher than those from the main experiment. More surprisingly, this easier setting benefits some models more than others. The relative performance of ClinicalBERT in particular is now clearly better, with this model achieving the best results for *Symptoms* $\rightarrow$ *Disease*. Furthermore, the standard BERT model now clearly underperforms the biomedical LMs, except for *Procedures* $\rightarrow$ *Disease* where it outperforms ClinicalBERT and BioBERT.

5 Conclusion

We have proposed DisKnE, a new benchmark for analysing the extent to which biomedical language models capture knowledge about diseases. Positive examples were obtained from MedNLI and MEDIQA-NLI, by manually identifying and categorizing hypotheses that express that the patient has some disease. Negative examples were selected to be similar to the target disease. To prevent shortcut learning, the hypotheses were canonicalized, such that models only get access to the name of the disease that is inferred. Our empirical analysis shows that existing biomedical language models particularly struggle with cases that require medical knowledge. The relative performance on the different categories suggests that different (biomedical) LMs have complementary strengths.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

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.

Vladimir Araujo, Andres Carvallo, Carlos Aspillaga, and Denis Parra. 2020. On adversarial examples for biomedical NLP tasks. *arXiv:2004.11157*.

Carlos Aspillaga, Andrés Carvallo, and Vladimir Araujo. 2020. Stress test evaluation of transformer-based models in natural language understanding tasks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1882–1894.

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. 2019. 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.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4762–4779.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7456–7463.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Joe Davison, Joshua Feldman, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1173–1178.

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*, pages 4171–4186.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*, pages 1753–1759.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 650–655.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Boran Hao, Henghui Zhu, and Ioannis Paschalidis. 2020. Enhancing clinical bert embedding using a biomedical knowledge base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. 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.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019b. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

William Kearns, Wilson Lau, and Jason Thomas. 2019. UW-BHI at MEDIQA 2019: An analysis of representation methods for medical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 500–509.

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*.

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.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Mingming Lu, Yu Fang, Fengqi Yan, and Maozhen Li. 2019. Incorporating domain knowledge into natural language inference on clinical texts. *IEEE Access*, 7:57623–57632.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alex Wong. 2020. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *arXiv:2010.10391*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534.

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. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing, pages 2463–2473.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5418–5426.

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, Brussels*, pages 1586–1596.

Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019. Incorporating domain knowledge into medical NLI using knowledge graphs. 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 2019*, pages 6091–6096.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics - on what language model pre-training captures. *Trans. Assoc. Comput. Linguistics*, 8:743–758.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Zhaofeng Wu, Yan Song, Sicong Huang, Yuanhe Tian, and Fei Xia. 2019. WTMED at MEDIQA 2019: A hybrid approach to biomedical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 415–426.