Journal of the American Medical Informatics Association, 28(9), 2021, 1970–1976
doi: 10.1093/jamia/ocab086
Advance Access Publication Date: 21 June 2021
Brief Communications

Brief Communications

Clinically relevant pretraining is all you need

Oliver J. Bear Don't Walk IV, Tony Sun, Adler Perotte, and Noémie Elhadad

Department of Biomedical Informatics, Columbia University, New York, New York, USA

Corresponding Author: Oliver J. Bear Don't Walk IV, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, PH20 3720, New York, NY 10032, USA; Ob2285@cumc.columbia.edu

Received 23 October 2020; Revised 19 April 2021; Editorial Decision 26 April 2021; Accepted 3 May 2021

ABSTRACT

Clinical notes present a wealth of information for applications in the clinical domain, but heterogeneity across clinical institutions and settings presents challenges for their processing. The clinical natural language processing field has made strides in overcoming domain heterogeneity, while pretrained deep learning models present opportunities to transfer knowledge from one task to another. Pretrained models have performed well when transferred to new tasks; however, it is not well understood if these models generalize across differences in institutions and settings within the clinical domain. We explore if institution or setting specific pretraining is necessary for pretrained models to perform well when transferred to new tasks. We find no significant performance difference between models pretrained across institutions and settings, indicating that clinically pretrained models transfer well across such boundaries. Given a clinically pretrained model, clinical natural language processing researchers may forgo the time-consuming pretraining step without a significant performance drop.

Key words: deep learning, natural language processing, transfer learning, social determinants of health, international classification of disease

INTRODUCTION

The electronic health record (EHR) contains a wealth of rich, unstructured patient health data, such as clinical text. Natural language processing (NLP) techniques allow for clinical text to be leveraged in a multitude of scenarios, such as information extraction,[1–3] understanding clinical workflow,[3,4] decision support,[5] and question answering.[6] NLP models often suffer reduced performance when applied across institutions[7,8] or specialties.[9] Drops in performance are due in part to differences in vocabulary, content, and style that manifest along axes such as syntax,[10–12] semantics,[13,14] and workflow procedures.[7] Historically, transferred NLP models overcome clinical institution differences by retraining models from scratch[7] or using domain adaptation techniques[15] for the downstream task of interest.

Recently, pretraining has led to robust methods for creating generalizable models that can be transferred to downstream tasks across genres and domains.[16–20] In contrast to traditional approaches, in which a new supervised task is learned from scratch on a training set, a pretrained model can leverage parameters that have already been trained to a different (simpler and often self-supervised) task. The intuition for this approach is that some of these parameters can generalize to the new task. In this way, previous experiences can be built upon. Furthermore, pretraining is inspired by the idea that certain features are learned across multiple tasks. For example, a model might learn how certain words relate to one another regardless of the task.[21] The fact that pretraining learns parameters that would otherwise need to be relearned by a new task makes pretraining especially useful when labeling data is time-consuming and expensive.

In traditional training zapproaches, training occurs in a single phase in which a model is initialized and trained on a task. This approach does not leverage shared information between tasks. In contrast, pretraining and transferring a model incurs a one-time cost of pretraining a model to extract generalizable features and transferring this same model to multiple tasks. Instead of relearning all parameters, pretrained parameters are updated to the specific task at hand. A visual depiction of the difference between approaches can be seen in Figure 1.

While pretraining has been commonly used to learn low level parameters like word embeddings,[22–24] recent advances have shown that pretraining is a powerful approach to learning higher level, generalizable linguistic representations.[16–20] Many kinds of approaches to pretraining have been tried, but language modeling tasks have proven to be generalizable to many other NLP tasks.[25] Specifically, masked language modeling, in which a model learns to identify masked words given the surrounding context has been a viable pretraining task.[16] Leveraging language modeling tasks, like masked language modeling, during pretraining is especially useful, as there are large amounts of text data to be leveraged that do not require any labeling. A more detailed explanation of language modeling can be found in the Supplementary Appendix A. Through language modeling-based pretraining, unlabeled data can be used to improve performance on a new task instead of the potentially costly step of labeling more task specific data.

For a variety of pretrained models, pretraining was initially performed using Wikipedia,[16,17,19,20] the Book Corpus,[16,18,20] the One Billion Word Benchmark,[19] news articles,[20] and Web snippets.[19,20] These corpora represent a more general domain and contain a wide variety of topics without specializing in any single topic. Pretraining on such corpora has worked well for many tasks in the general domain; however, the clinical domain containsspecializations like language, abbreviations, grammar, and semantics are not encountered in the general domain, leaving room for domain-specific pretraining. This has led to clinical domain variants of pretrained models,[26–28] which have outperformed their general domain counterparts on a variety of clinical NLP tasks such as readmission prediction,[27] named entity recognition,[26,28] reason for visit extraction,[29] natural language entailment,[26,28] and medication extraction.[30]

Beyond differences between domains, heterogeneity within the clinical domain such as geography, clinical setting, patient population, and de-identification status manifest along multiple axes such as syntax,[10–12] semantics,[13,14] and workflow procedures.[7] It is generally accepted that NLP model performance may degrade when evaluated on data with a different distribution than what had been trained on and is nontrivial to deal with.[7,15] Many pretrained models in the clinical domain available for download are pretrained using the Medical Information Mart for Intensive Care-III(MIMIC-III) dataset.[31] Given the differences between clinical institutions and settings, we ask the following questions. Is a single round of clinically relevant pretraining sufficient to generalize across multiple clinical institutions and settings? Furthermore, can institution- or setting-specific pretraining improve downstream task performance over pretraining at a different institution or setting? These questions are relevant for clinical NLP researchers looking to apply clinically relevant pretrained models to their own data, in which pretraining their own model might be prohibitively expensive. Using a meticulous experimental design, we explore whether institutional differences impact

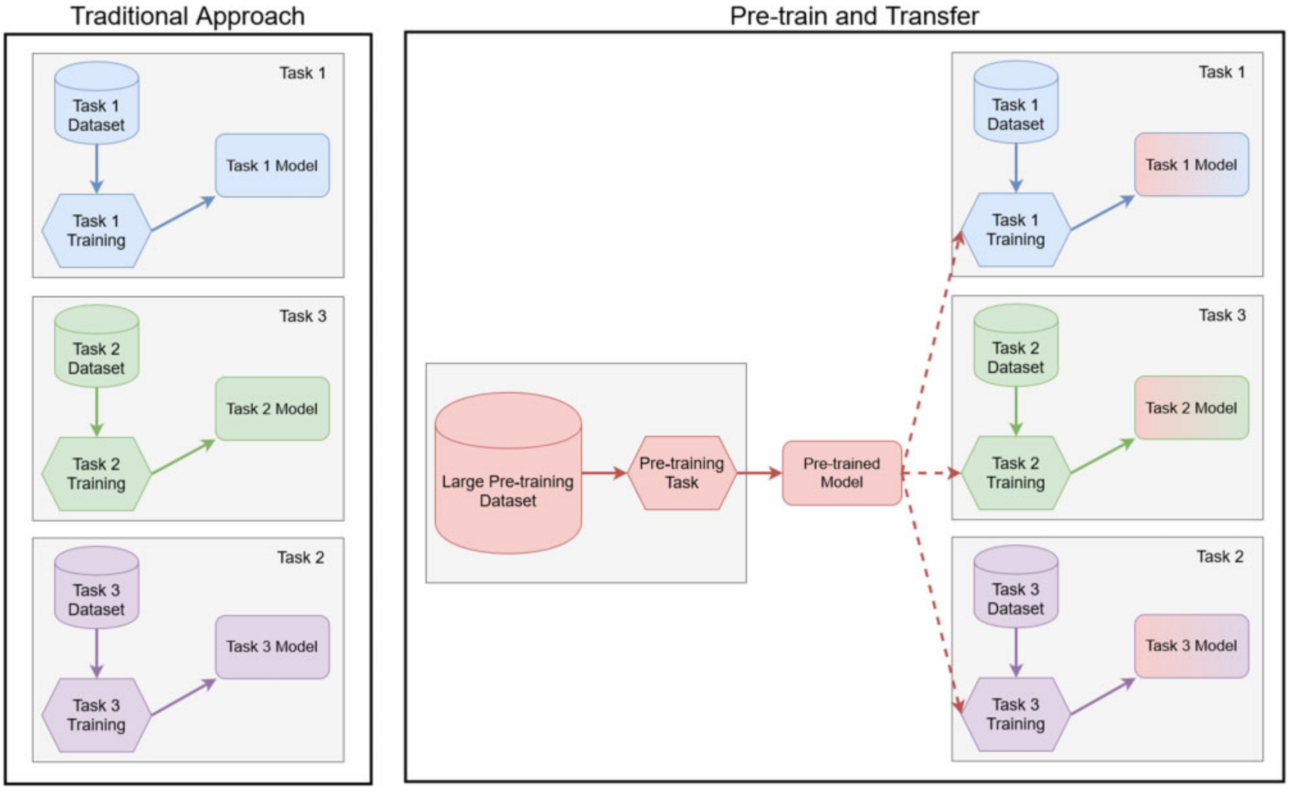

Figure 1. An overview of the pretraining and transfer phases vs a traditional training approach. The traditional training approach initializes a new model for each task without sharing knowledge between tasks. In contrast, during the pretraining phase a model learns parameters that can generalize to other natural language processing tasks by learning a pretraining task. Pretraining datasets can be large, allowing tasks with smaller datasets to take advantage of the "warm start" provided through pretraining. Pretraining is a one-time cost, allowing for a pretrained model to be transferred to multiple new tasks. During the transfer phase, the pretrained model is updated to perform a new task and can result in better performance with less data than if the model was randomly initialized.

performance on downstream tasks when pretraining at the same or a different institution as the downstream task. Our results indicate that institution- or setting-specific pretraining does not meaningfully improve performance and clinically relevant pretraining is all you need.

MATERIALS AND METHODS

Using EHR data from 2 institutions, we assess the impact of pretraining on different institutional and setting data on our downstream document classification tasks. We collect 1 general and 2 intensive care unit (ICU) corpora from the 2 institutions and 2 downstream task datasets from each institution. Using the 3 pretraining corpora, we create 3 pretrained models, and evaluate the performance of each model trained on each downstream task. Here, training refers to updating a model's weights from those learned during pretraining to a new task. In this work we focus on using the Bidirectional Encoder Representations from Transformers (BERT)[16] model, as it has been shown to be a strong baseline for state-of-the-art pretrained models. Models are pretrained using the pretraining tasks outlined in the original BERT article.[16] A more detailed explanation of the pretraining methods used in this work can be found in the Supplementary Appendix A. The proposed experimental design, outlined in Figure 2A and 2B, allows us to measure the impact of the pretraining and downstream task data come from the same institution or setting on. We measure impact using downstream task performance.

Datasets for pretraining

We leverage 3 corpora from 2 different institutions. The first institutional data is a collection of ICU clinical data from Beth Israel Deaconess Medical Center[31] between 2001 and 2012 (MIMIC). The second institutional data is from Columbia University Irving Medical Center (CUIMC) between 2005 and 2015.

One general clinical corpus and 2 ICU specific corpora were generated. GEN-C is a random selection of notes from CUIMC without any specification for setting. ICU-M is a random selection of notes from the ICU-specific dataset MIMIC, while ICU-C is a random selection of ICU notes from CUIMC. In this work, we control for the number of tokens and training examples in each corpus. The number of tokens in each corpus is used as a statistic to represent how much BERT has been pretrained according to the original authors of BERT.[16] In order to avoid data leakage, any data from the test sets of the downstream task were removed from the pretraining corpora. Pretraining demographic (raceand ethnicity data have been merged between MIMIC and CUIMC which collected this information differently; specifically, the Hispanic or Latinx category for Gen-C and ICU-C is not mutually exclusive from the ethnic categories, while ICU-M kept this category mutually exclusive) and text information can be found in Tables 1 and 2. More information about the pretraining data can be found in Supplementary Appendix B.

The 3 corpora, GEN-C, ICU-M, and ICU-C, allow for model performance comparisons on downstream tasks while either varying or holding constant the pretraining setting or institution. Models pretrained using GEN-C and ICU-C data are examples of pretrained models at the same institution but different settings. While models pretrained using ICU-C and ICU-M are examples of models pretrained with same setting but different institutions. Finally, models pretrained using GEN-C and ICU-M are examples of different institutions and different settings. All 3 of these scenarios provide insight into the importance of pretraining models at the same institution or setting.

Datasets for downstream tasks

Two multilabel document classification tasks were chosen to train and evaluate on. International Classification of Diseases–Ninth Revision (ICD-9) code classification was chosen, as ICD-9 codes are prevalent across many institutions and represents tasks with large but biased datasets. Social determinants of health (SDH) classification was chosen as a representative task for scenarios with smaller datasets. More importantly, these 2 downstream tasks were chosen as they can vary greatly between institutions or settings. SDH have been shown to have high lexical variation even when discussing the same concepts and can vary greatly by population and geography.[32,33] For example, homelessness can be indicated by naming a local homeless shelter in which a patient resides, which would not be a readily identifiable indicator for homelessness at other institutions. Furthermore, an SDH task is a practical example for pretraining because labeling SDH is time-consuming and requires expert knowledge. ICD-9 code distribution can also vary by setting, as was shown in our datasets in which the top 50 ICD-9 codes for MIMIC (ICU setting) and CUIMC (setting agnostic) only shared 24 codes. Beyond these reasons, access to these datasets at both intuitions made the experimental design possible.

Each institution, MIMIC and CUIMC, have training, validation, and test sets for both downstream tasks. ICD codes are extracted from the EHR at each institution and matched to clinical notes. We limit ourselves to classifying the top 50 ICD-9 codes at each institution but do not remove notes without any of these codes. The SDH classification corpora are annotated at a document level with 5 SDH categories: smoking status, illicit drug use status, housing status, sexuality documented;, and sexual history documented. Further details provided in previous work.[32] All training, validation, and test splits are made at the patient level to avoid data leakage and each model uses the same training, validation, and test sets. Table 3 summarizes dataset sizes, while Table 4 summarizes dataset demographics (race and ethnicity data have been merged between MIMIC and CUIMC, which collected this information differently; specifically, the Hispanic or Latinx category for CUIMC data is not mutually exclusive from the ethnic categories, while MIMIC kept this category mutually exclusive). More information about the distribution of ICD and Social and Behavioral Determinants of Health (SBDH)codes can be found in Supplementary Appendix C.

Experimental design for pretraining

The BERT pretrained model,[16] which is not specialized on clinical or biomedical data, and the PubMedBERT pretrained model,[28] which consists of BERT further pretrained on biomedical articles, are the 2 baselines pretrained models for our experiments. Both models were pretrained using the same tasks as the current work but relied on different pretraining data. Practically, further pretraining consists of another round of pretraining, following the BERT procedures.

Starting from PubMedBERT, we further pretrain 3 different pretrained models: BERT-IM leveraging ICU-M, BERT-GC leveraging GEN-C, and BERT-IC leveraging ICU-C. BERT models further pretrained with biomedical data have been shown to outperform BERT on clinical datasets,[26,28] and PubMed presents a much larger dataset than any single clinical dataset, thus making PubMedBERT an ideal initialization for clinically relevant pretraining.

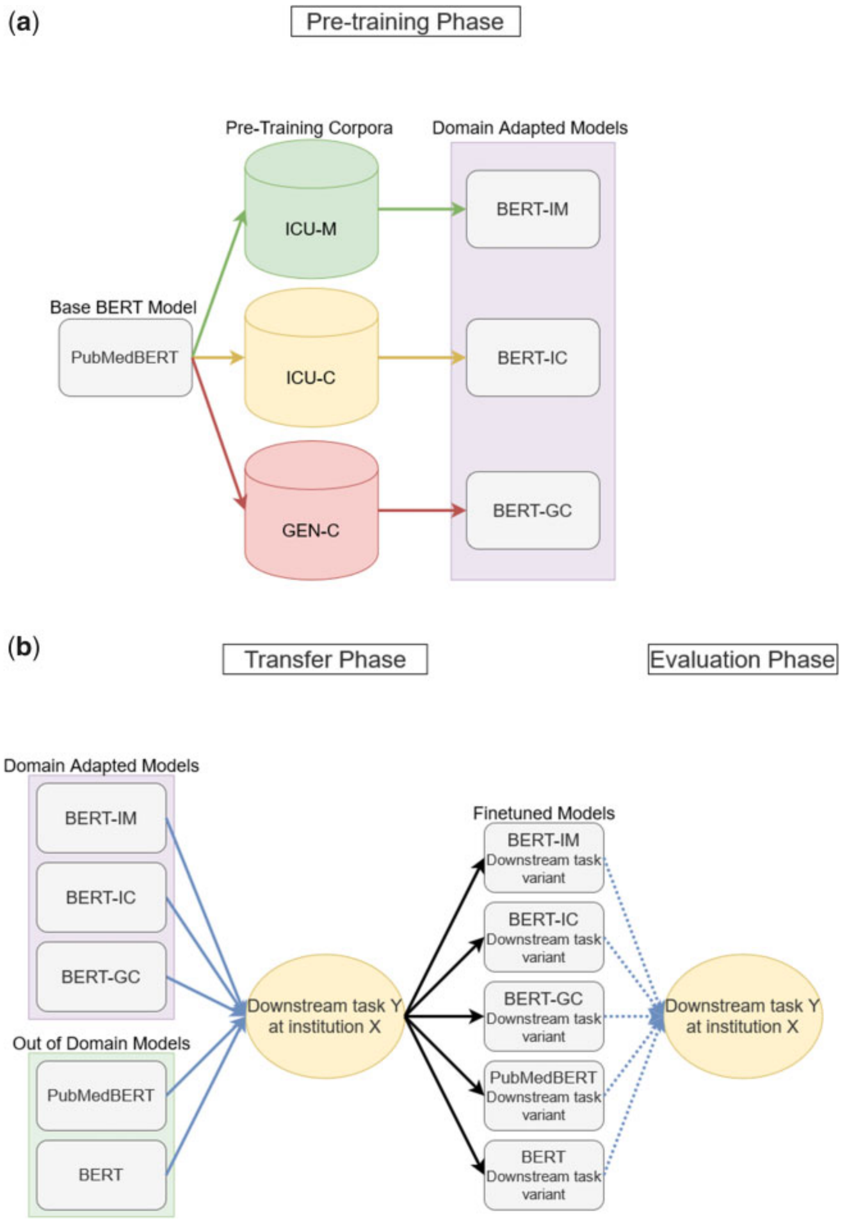

Figure 2. (A) Experimental design for pretraining. Further pretraining is performed on PubMedBERT using 3 corpora to create 3 new models. (B) Clinical and non-clinical Bidirectional Encoder Representations from Transformers (BERT) models are transferred and then evaluated on downstream tasks at each institution.

Table 1. Summary statistics for the pretraining corpora

	ICU-M	GEN-C	ICU-C
Patients	35 000	109 000	32 000
Notes	134 000	280 000	148 000
Tokens	254 000 000	255 000 000	255 000 000
Examples	260 000 000	255 000 000	255 000 000

Number of examples is calculated as the number of individual text chunks (ie, observations) in the pretraining dataset.

Table 2. Pretraining corpora patient demographics

	ICU-M	GEN-C	ICU-C
White	24 692	27 660	12 728
Black/African American	2888	8110	3214
Hispanic or Latinx	1261	24 031	15 307
Asian	1163	1240	678
Native American/Alaskan Native	34	97	85
Native Hawaiian or Pacific Islander	11	349	86
Unknown/not specified	5281	71 709	15 307

We used a learning rate of 1×10^{-4}, a linear warm up schedule of 10% of the total number of steps, a batch size of 500. Finally, each observation was a maximum length of 128 tokens. Following the original pretraining data generation in Devlin et al,[16] we concatenated nonoverlapping sentences up to 128 tokens in length. Masking was carried out following the masking procedure of the original BERT article[16] by masking, replacing, or leaving a token unchanged. These observations consisted of 2 segments, in which 50% of the time the second segment followed the first segment in

Table 3. Downstream task observation splits

CUIMC						MIMIC					
ICD			SDH			ICD			SDH		
Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
28 000	14 000	12 000	3000	625	689	24 000	1600	3.3 000	91	37	56

CUIMC: Columbia University Irving Medical Center; ICD: International Classification of Diseases; MIMIC: Medical Information Mart for Intensive Care; SDH: social determinants of health.

Table 4. Downstream task race and ethnicity splits

	CUIMC		MIMC	
	ICD	SDH	ICD	SDH
White	9064	110	14 503	44
Black/African American	2869	190	1740	11
Hispanic or Latinx	6412	239	699	3
Asian	325	3	533	0
Native American/Alaskan Native	28	0	19	0
Native Hawaiian or Pacific Islander	64	3	7	0
Unknown/not specified	11 130	647	2893	40

CUIMC: Columbia University Irving Medical Center; ICD: International Classification of Diseases; MIMIC: Medical Information Mart for Intensive Care; SDH: social determinants of health.

the document and the other 50% of the time the second segment was randomly selected from the corpus. The 2 segments were used for the additional pretraining task next sentence prediction. Learning was performed on 1 NVIDIA GeForce RTX 2080 Ti GPU using PyTorch with mixed precision. Each model pretrained for 10 epochs over approximately 2.5 days.

Experimental design for downstream tasks

We train and test 2 downstream tasks, ICD and SDH document-level classification. For each task, we want to assess whether aligning the pretraining data and the data used for training the task itself (either by institution or setting) benefit its performance.

Given a task and institution, for instance ICD classification and CUIMC, we control for training, validation, and testing sets and compare performance on the task when using different pretrained models. As such, for each corpus and task, there are 5 models that are trained, validated, and tested. We use the validation set to tune the maximum number of epochs (3, 4, 10) used during downstream task training.

Because both tasks operate at the document level, and because clinical notes are particularly long documents, each clinical note is broken down into up to 10, nonoverlapping, 128-token chunks (n). Following the approach of Huang et al,[27] the probability of a document's classification into category k is based on the n classified chunks. Rather than computing an average probability over the n chunks for category k, it also takes into consideration the maximum probability over all chunks using a combination of average and max pooling. Letting $\mathrm{P}^{\mathrm{n}}_{\mathrm{mean}}$ and $\mathrm{P}^{\mathrm{n}}_{\mathrm{max}}$ be the mean and max probability over all n chunks, respectively, the final probability is $\mathrm{P}(k=1)=\frac{\mathrm{P}^{\mathrm{n}}_{\mathrm{max}}+\mathrm{P}^{\mathrm{n}}_{\mathrm{mean}}\frac{\mathrm{n}}{2}}{1+\frac{\mathrm{n}}{2}}$.

All results are measured in macro-averaged average precision. Average precision summarizes the precision-recall curve by summing the precision at different thresholds weighted by the change in recall from the previous threshold.[34] Bootstrapped performances and 95% confidence intervals are calculated by evaluating all models on 1000 bootstrapped test sets. For a given downstream task all models are trained on the same training set and evaluated on the same 1000 bootstrapped test sets. All results are presented using the bootstrapped average performance and 95% confidence intervals.

RESULTS

All clinical models outperform the baseline models on the 2 downstream tasks of ICD and SDH classification and across institutions. Similarly, PubMedBERT outperforms the original BERT on both downstream tasks and across institutions. We note however that on the SDH downstream task, the confidence intervals of PubMedBERT and BERT overlap.

For the ICD classification downstream task, we first note that a model's performance on MIMIC ICD is better than its performance on CUIMC ICD across all models. This is not surprising: while the MIMIC dataset contains only ICU admissions, the CUIMC dataset is more heterogeneous with different settings, leading to higher-perplexity tasks.

We also note that, as expected, the range of the confidence intervals for the different models across tasks and institutions is directly related to the size of training and testing data. That is, the SDH tasks, especially MIMIC SDH, have larger and possibly overlapping confidence intervals due to how small these datasets are compared with the easier and cheaper-to-label ICD datasets.

Of interest to our original research question, we see that there are slight differences in performance from one setting to the next and from one institution to the next. On balance, taking their confidence intervals into account, all clinical pretrained models yield similar performance to each other across all tasks. These results are summarized in Figures 3 and 4.

DISCUSSION

The pretrain and transfer paradigm in NLP has led to an explosion of domain-specific models that have achieved state-of-the-art performance across many tasks. In this work, we explored how well pretrained BERT models transfer across institutional and setting boundaries. We confirm previous results that as BERT is pretrained on data closer to the clinical domain, model performance improves. The clinically adapted BERT variants outperform nonclinical BERT models in 3 of 4 experiments, in which in the fourth experiment the performances are tied. Overall, clinical BERT models perform similarly across institutional and setting boundaries regardless of the pretraining setting or institution.

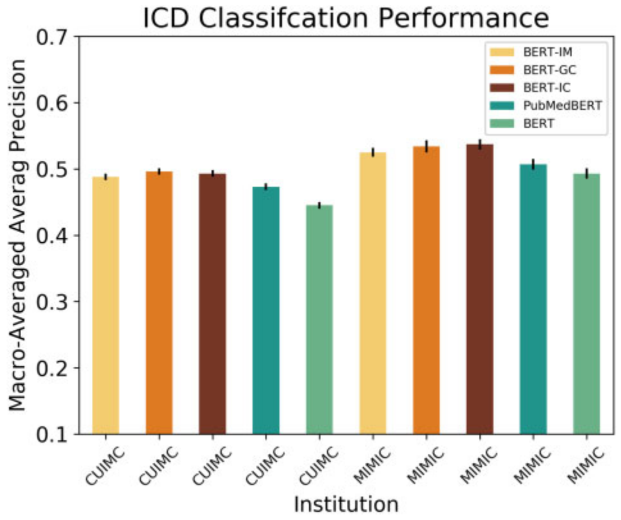

Figure 3. Macro-averaged average precision for International Classification of Diseases (ICD) classification across institutions. The ICU-M, GEN-C, and ICU-C corpora are used to pretrain BERT-IM, BERT-GC, and BERT-IC. PubMedBERT and Bidirectional Encoder Representations from Transformers (BERT) are baseline models. CUIMC: Columbia University Irving Medical Center; MIMIC: Medical Information Mart for Intensive Care.

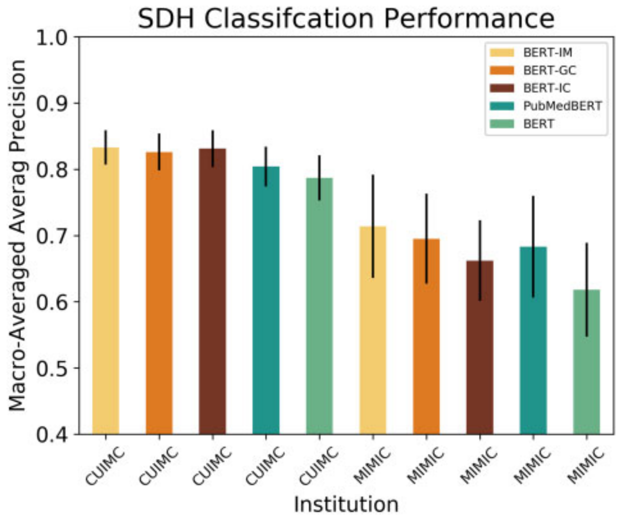

Figure 4. Macro-averaged average precision for social determinants of health (SDH) classification across institutions. The ICU-M, GEN-C, and ICU-C corpora are used to pretrain BERT-IM, BERT-GC, and BERT-IC. PubMedBERT and Bidirectional Encoder Representations from Transformers (BERT) are baseline models. CUIMC: Columbia University Irving Medical Center; MIMIC: Medical Information Mart for Intensive Care.

To answer the question of whether institution-specific pretraining is helpful, we conclude that there is no statistical difference between clinical BERT variants. There is evidence of a small differences, specifically BERT-IC and BERT-GC on MIMIC-ICD, in which BERT-IC outperforms BERT-GC, while this result is reversed on CUIMC-ICD. This could be evidence of the importance of matching the setting when transferring models to new institutions. However, this difference, and others, are small enough as not to be considered meaningfully different.

While testing all available clinical BERT models might provide some performance improvement, there is no guarantee of a statistically significant performance increase even if the downstream and pretraining data match across institution or setting. These results raise the question of whether the investment into setting or institution pretraining is warranted. We note that the results presented here are not at odds with the practice of adapting specific NLP task models to new institutions or settings. While it may not be necessary to adapt pretrained models to new institutions or settings at the level of pretraining, it is likely still necessary to adapt such models when they have been specialized to a specific NLP task. It should be noted that this work only explores 2 document classification tasks. There might also be downstream tasks on specialized corpora in which further pretraining does confer a meaningful improvement. In the future, we plan to explore entity-level classification tasks, and performance on the BLUE dataset,[28] though we cannot perform a bidirectional comparison in this case without parallel datasets.

FUNDING

This work was supported by a National Library of Medicine award T15 LM0007079 and a National Institute of General Medical Sciences award R01 GM114355.

AUTHOR CONTRIBUTIONS

OBDW, NE, and AP were involved in study planning. OBDW was involved in modeling. OBDW and TS were involved in data curation. All authors contributed to writing and reading and have approved the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

DATA AVAILABILITY STATEMENT

The data underlying this article cannot be shared publicly due to protected health information. The data will be shared on a reasonable request to the corresponding author.

REFERENCES

1. Flynn RWV, Macdonald TM, Schembri N, *et al.* Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. *Pharmacoepidemiol Drug Saf* 2010; 19: 843–7.
2. Yang H, Spasic I, Keane JA, *et al.* A text mining approach to the prediction of disease status from clinical discharge summaries. *J Am Med Inform Assoc* 2009; 16: 596–600.
3. Friedman C, Alderson PO, Austin JH, *et al.* A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994; 1: 161–74.
4. Ou Y, Patrick J. Automatic structured reporting from narrative cancer pathology reports. *Electron J Health Inform* 2014; 8: .
5. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol* 2014; 12 (7): 1130–6.
6. Ben Abacha A, Zweigenbaum P. MEANS: a medical question-answering system combining NLP techniques and semantic Web technologies. *Inf Process Manag* 2015; 51 (5): 570–94.

7. Sohn S, Wang Y, Wi C-I, *et al.* Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *J Am Med Inform Assoc* 2018; 25: 353–9.
8. Liu M, Shah A, Jiang M, *et al.* A study of transportability of an existing smoking status detection module across institutions. *AMIA Annu Symp Proc* 2012; 2012: 577–86.
9. Bakken S, Hyun S, Friedman C, *et al.* A comparison of semantic categories of the ISO reference terminology models for nursing and the MedLEE natural language processing system. *Stud Health Technol Inform* 2004; 107: 472–6.
10. Stetson PD, Johnson SB, Scotch M, *et al.* The sublanguage of cross-coverage. *Proc AMIA Symp* 2002; 742–6.
11. Friedman C. A broad-coverage natural language processing system. *Proc AMIA Symp* 2000; 270–4.
12. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform* 2002; 35 (4): 222–35.
13. Xu H, Stetson PD, Friedman C. Methods for building sense inventories of abbreviations in clinical notes. *J Am Med Inform Assoc* 2009; 16: 103–8.
14. Wu Y, Denny JC, Trent Rosenbloom S, *et al.* A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). *J Am Med Inform Assoc* 2017; 24: e79–86.
15. Zhang Y, Tang B, Jiang M, *et al.* Domain adaptation for semantic role labeling of clinical text. *J Am Med Inform Assoc* 2015; 22 (5): 967–79.
16. Devlin J, Chang M-W, Lee K, *et al.* 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)*. Stroudsburg, PA: Association for Computational Linguistics; 2019: 4171–86.
17. Howard J, Ruder S. Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA: Association for Computational Linguistics; 2018: 328–39.
18. Radford A, Narasimhan K, Salimans T, *et al.* Improving language understanding by generative pre-training. 2018. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf Accessed April, 5, 2020.
19. Peters M, Neumann M, Iyyer M, *et al.* Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers*). Stroudsburg, PA: Association for Computational Linguistics; 2018: 2227–37.
20. Liu Y, Ott M, Goyal N, *et al.* RoBERTa: a robustly optimized BERT pretraining approach. arXiv, http://arxiv.org/abs/1907.11692, 26 Jul 2019, preprint: not peer reviewed.
21. Mou L, Meng Z, Yan R, *et al.* How transferable are neural networks in NLP applications? In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics; 2016: 479–89.
22. Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems–Volume 2*. New York, NY: Association for Computing Machinery; 2013: 3111–9.
23. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA: Association for Computational Linguistics; 2014: 1532–43.
24. Joulin A, Grave E, Bojanowski P, *et al.* Bag of tricks for efficient text classification. arXiv, http://arxiv.org/abs/1607.01759, 6 Jul 2016, preprint: not peer reviewed.
25. Wang A, Hula J, Xia P, *et al.* Can you tell me how to get past Sesame Street? Sentence-level pretraining beyond language modeling. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics; 2019: 4465–76.
26. Alsentzer E, Murphy J, Boag W, *et al.* Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Stroudsburg, PA: Association for Computational Linguistics; 2019: 72–8.
27. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv, http://arxiv.org/abs/1904.05342, 11 Ap 2019, preprint: not peer reviewed.
28. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on 10 benchmarking datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Stroudsburg, PA: Association for Computational Linguistics; 2019: 58–65.
29. Valmianski I, Goodwin C, Finn IM, *et al.* Evaluating robustness of language models for chief complaint extraction from patient-generated text. arXiv, http://arxiv.org/abs/1911.06915, 15 Nov 2019, preprint: not peer reviewed.
30. Selvaraj SP, Konam S. Medication regimen extraction from medical conversations. arXiv, http://arxiv.org/abs/1912.04961, 10 Dec 2019, preprint: not peer reviewed.
31. Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3: 160035.
32. Feller DJ, Zucker J, Bear DWIO, *et al.* Towards the inference of social and behavioral determinants of sexual health: Development of a gold-standard corpus with semi-supervised learning. *AMIA Annu Symp Proc* 2018; 2018: 422–9.
33. Bejan CA, Angiolillo J, Conway D, *et al.* Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc* 2018; 25: 61–71.
34. Su W, Yuan Y, Zhu M. A relationship between the average precision and the area under the ROC curve. In: *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*. New York, NY: Association for Computing Machinery; 2015: 349–52.